DOCUMENT RESUME

ED 440 148 TM 030 760

AUTHOR Koretz, Daniel; Hamilton, Laura
TITLE Assessing Students with Disabilities in Kentucky: The Effects of Accommodations, Format, and Subject. CSE Technical Report.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-498
PUB DATE 1999-01-00
NOTE 63p.
CONTRACT R305B60002
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Disabilities; Elementary Secondary Education; *Item Bias; Scores; Test Format; Test Use
IDENTIFIERS *Kentucky; *Testing Accommodations (Disabilities)

ABSTRACT

An earlier study (D. Koretz, 1997) found that Kentucky had been unusually successful in testing most students with disabilities, but it also found numerous signs of poor measurement, including differential item functioning (DIF) in mathematics, apparently excessive use of accommodations, and implausibly high mean scores for some groups of students with disabilities. This study used newer data to test the stability of the findings over a 2-year period, to extend some of the analyses to additional subject areas, and to compare performance on open-response items to that on multiple-choice items. The researchers analyzed data from students in grades 4, 5, 7, 8, and 11. The inclusiveness of the assessment persisted, and the frequency of special accommodations remained unchanged. The mean performance of elementary school students with disabilities dropped substantially, apparently because of a lessened impact of accommodations on scores. These lower scores appear to be more plausible. The differences in scores between disabled and nondisabled students tended to be larger on multiple-choice components in the early grades. Across grades, the effects of accommodations were stronger on the open-response items. Results provide some evidence that score accommodations change the dimensionality of the assessment. DIF was found in both test formats. Further research and more detailed data are needed to clarify the reasons for the findings. (Contains 19 tables, 15 figures, and 10 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

ED 440 148

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Assessing Students With Disabilities in Kentucky: The Effects of Accommodations, Format, and Subject

CSE Technical Report 498

Daniel Koretz
CRESST/RAND Education

Laura Hamilton
RAND Education

TM030760

UCLA Center for the Study of Evaluation

In Collaboration With

BEST COPY AVAILABLE

Assessing Students With Disabilities in Kentucky: The Effects of Accommodations, Format, and Subject

CSE Technical Report 498

Daniel Koretz
CRESST/RAND Education

Laura Hamilton
RAND Education

January 1999

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.3 Accommodation Daniel Koretz, Project Director, RAND Education

Copyright © 1999 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

ASSESSING STUDENTS WITH DISABILITIES IN KENTUCKY: THE EFFECTS OF ACCOMMODATIONS, FORMAT, AND SUBJECT

Daniel Koretz, CRESST/RAND Education

Laura Hamilton, RAND Education

Abstract

In an earlier study (Koretz, 1997), we reported that Kentucky had been unusually successful in testing most students with disabilities, but we found numerous signs of poor measurement, including differential item functioning in mathematics, apparently excessive use of accommodations, and implausibly high mean scores for some groups of students with disabilities. This study used newer data to test the stability of the findings over a two-year period, to extend some of the analyses to additional subject areas, and to compare performance on open-response items to that on multiple-choice items, which were not administered in the assessment investigated earlier. We analyzed test score data from students in Grades 4, 5, 7, 8, and 11.

The inclusiveness of the assessment persisted, and the frequency of specific accommodations remained unchanged. The mean performance of elementary school students with disabilities dropped substantially, however, apparently because of a lessened impact of accommodations—particularly dictation—on scores. These lower scores, while discouraging as an indication of student performance, appear to be more plausible. The differences in scores between disabled and non-disabled students tended to be larger on the multiple-choice components in the elementary grades, whereas the differences were generally similar or larger for open-response components in the higher grades. Across grades, the effects of accommodations were stronger on the open-response than on the multiple-choice tests.

Correlations among parts of the assessment were different for accommodated students with disabilities than for others, with higher correlations across subjects for the open-response components. This may indicate that some accommodations change the dimensionality of the assessment. DIF was apparent in both the open-response and multiple-choice components of the assessment, but it was mostly limited to students who received accommodations. DIF was found in both formats. Further research and more detailed data concerning the specific uses of accommodations are needed to clarify the reasons for these findings and to guide the development of more effective approaches to the assessment of students with disabilities.

Background

In recent years, policy changes at the national and state levels have pressed for increased inclusion of students with disabilities in large-scale assessments, including both those used for monitoring, such as the National Assessment of Educational Progress (NAEP), and those used for accountability. This change has come at a time when the use of formats other than multiple choice—such as open-response paper-and-pencil tasks and hands-on performance tasks—has become increasingly routine in large-scale assessments.

Efforts to increase the participation of students with disabilities in large-scale assessments, however, are hindered by a lack of experience and systematic information (National Research Council, 1997). For example, there is little systematic information on the use or effects of special testing accommodations for elementary and secondary students with disabilities. In addition, there is little evidence about the effects of format differences on the assessment of students with disabilities. Some observers have argued that the use of formats other than multiple choice will make assessments fairer for many students, including some with disabilities. Others have argued the opposite, pointing out that open-response questions, for example, mix verbal skills with other skills to be measured and may make it more difficult to isolate and compensate for the effects of disabilities. Relevant research, however, is scarce.

The statewide assessment program in Kentucky, the Kentucky Instructional Results Information System (KIRIS), provides valuable insight into these questions. Kentucky is one of the most ambitious states in its efforts to include most students with disabilities in its regular statewide assessment, and proctors provide information on students' primary disabilities and assessment accommodations. In addition, KIRIS provides a comparison between multiple-choice and open-response paper-and-pencil formats. For several years, KIRIS included no multiple-choice items, but they were gradually reintroduced over the past several years in response to criticism that their absence limited content coverage and impeded linking of scores over time (Hambleton et al., 1995).

In a previous study using 1995 data that included no multiple-choice items, we studied the participation of students with disabilities in KIRIS and the quality

of the performance data they generated (Koretz, 1997).[1] This earlier study found that Kentucky had succeeded in including most of its students with disabilities in the regular KIRIS assessment. It found little evidence of differential item functioning (DIF) for disabled students assessed without accommodations and no evidence that items differentiated less well for students with disabilities than for others, regardless of the use of accommodations. On the other hand, it found several possible problems with the use of accommodations, including apparently excessive use of certain accommodations, implausibly high scores for students assessed with certain accommodations, and considerable DIF for the majority of disabled students who were assessed with accommodations. In addition, the study found that parts of the KIRIS assessment were too difficult for many students with disabilities.

This earlier study, however, included only a single test format: open-response pencil-and-paper questions. This is an important limitation. Some observers have noted that performance assessments may pose particular problems for some students with disabilities (e.g., National Research Council, 1997). This study using more recent (1997) data was therefore undertaken in part to compare the performance of students with disabilities on the multiple-choice and open-response portions of the KIRIS assessment. For example, it compared the performance correlates of accommodations and the incidence of DIF across the two formats, separately for mathematics, reading, science, and social studies in Grades 4, 8, and 11. This study also examined the stability of the earlier findings over time and extended some of the analyses that had been limited to reading and mathematics to science and social studies.

Apart from the inclusion of multiple-choice items, the 1997 assessment differed from the 1995 assessment in an important respect that should be borne in mind in comparing findings from the two years, In 1995, the core subjects considered here—reading, mathematics, science, and social studies—were all assessed in Grades 4, 8, and 11. In 1997, the elementary school and middle school assessments were each split between two grades. Reading and science were assessed in Grades 4 and 7, and mathematics and social studies were assessed in Grades 5 and 11. Most of the results presented here were similar for the two

---

[1] In this report, school years are identified by the date of testing, which was done in the spring. Thus, the 1994-95 school year is identified as 1995, and the 1996-97 school year is identified as 1997.

grades within each level, so we discuss some results as "elementary" or "middle school," without reference to specific grades, for simplicity. However, many of the tables show each grade separately, and we make reference to specific grades when it increases clarity.

The analyses reported here include only students who had scores for both the multiple-choice and open-response components of the assessment so that differences in performance on the two components would not be confounded with differences in the ability or other characteristics of students who had scores on the two components. The number of students excluded as a result was very small.

Kentucky's Policies for the Assessment of Students with Disabilities

Kentucky's policies for the assessment of students with disabilities remained largely unchanged from 1995 to 1997, so we repeat here the short description we provided in the earlier study (Koretz, 1997) for readers who are unfamiliar with the KIRIS system. We are aware of only one change that warrants mention. While the policies pertaining to the use of accommodations were not altered between 1995 and 1997, the implementation of them at the local level may have changed. There were some well-publicized allegations of inappropriate use of accommodations, and while most turned out to be unsubstantiated, this publicity may have altered practice (S. Trimble, personal communication, May 13, 1998). Such changes in practice, if they were sufficiently widespread, might help explain inconsistencies between the findings of this study and those of the earlier one.

Kentucky's policies for the assessment of students with disabilities are guided by the premise that only a small number of students with disabilities—1% to 2% of the total student population, comprising primarily students with moderate to severe cognitive disabilities—should be excluded from the regular KIRIS assessment. Most of those excluded are to be tested with a different assessment, called the KIRIS Alternate Portfolio. (Data from the Alternate Portfolio program are not considered in this report.) Students who are in ungraded programs are tested on the basis of age. The decision rules for determining inclusion are as follows:

- Students without an IEP or Section 504 plan participate in KIRIS without accommodation or modification.

- Students who meet several criteria indicating severe cognitive limitations, including being unable to complete a regular diploma program by reason of disability, even with extended services, accommodations, and modifications, are eligible for the KIRIS Alternative Portfolio.
- All students with IEPs or Section 504 plans who do not meet the preceding criterion are to be assessed using the regular KIRIS assessment.

Students with disabilities with IEPs or Section 504 plans may be administered KIRIS with either accommodations or modifications, subject to explicit limitations. (Kentucky defines an accommodation as "an alteration in the testing environment or process" and a modification as "an alteration in the assessment instrument" [Kentucky Department of Education, 1996, *Procedures for Considering Student Inclusion*, footnote 2].) State policy allows the use of "adaptations and modifications including the use of assistive technology devices that are consistent with the instructional strategies specified on the student's . . . IEP or 504 plan and available to the student in the course of his/her instructional process" (Program Advisory No. OCAA-93-94, February 9, 1993, cited in Kentucky Department of Education, 1996, Attachment G). These accommodations and modifications:

1. must be part of the student's ongoing instructional program;
2. may not be introduced for the first time during the KIRIS assessment;
3. must be "based on the individual needs of the students and not on a disability category"; and
4. shall not "inappropriately impact the content being measured." (Kentucky Department of Education, 1996, Attachment G, A1)

Accommodations Offered

Several of the accommodations commonly offered in some other assessment programs are not specifically recorded in the KIRIS assessment. Provision of additional time is one of the most commonly offered accommodations in some assessment programs. Most parts of KIRIS, however, are not intended to be speeded, and additional time can be offered to both disabled and other students without any notation on the testing record. Kentucky students with disabilities might in fact be offered additional time more frequently than other students, or might be offered on average more additional

time, but there are presently no data pertaining to this question. Kentucky Department of Education (KDE) guidelines indicate that it is permissible to provide students with disabilities with breaks during testing time, if doing so is consistent with their IEPs or 504 plans, but data pertaining to this accommodation are not collected. KDE neither provides guidance nor collects data about the use of separate assessment settings—another frequently offered accommodation. KDE makes KIRIS available in large-type and Braille formats. Oral presentation by tape is not available.

KDE collects information about six specific accommodations:

- paraphrasing;
- oral presentation of the assessment (providing a reader);
- allowing dictation of responses (providing a scribe);
- cueing;
- use of an interpreter; and
- technological aids.

In addition, proctors could indicate the use of other, unspecified accommodations.

Restrictions on the Use of Accommodations

KDE provides detailed guidelines about the use of accommodations in the KIRIS assessment, including numerous specific questions and answers about the uses of specific accommodations. Given the frequency with which various accommodations were used in KIRIS in 1995, the guidelines pertaining to paraphrasing, oral presentation, and dictation are particularly important.

Guidelines about the use of paraphrasing are specific and restrictive. The guidelines note that paraphrasing is allowed only for directions, not for reading and content passages. Paraphrasing is labeled an intrusive technique, and educators are told that they should use the least intrusive method possible. Paraphrasing can include repeating, rephrasing, or breaking down directions. However, it should not entail changes in "critical words" and should not be used "simply because vocabulary or content has not been taught/learned" (Kentucky Department of Education, 1996, Attachment G, A18). No concrete examples of appropriate or inappropriate paraphrasing are provided.

Guidelines for oral administration differentiate between the reading assessment and other content areas and include the following:

- On-demand tasks in general may be read to a student if the student has a verified disability in the area of reading, the student's IEP documents the use of a reader in instruction, and use of a reader "is not a replacement for reading instruction or technology" (Kentucky Department of Education, 1996, Attachment G, A20).
- Reading assessments "may be read to a student on the premise that the intent of reading is to measure comprehension, only if this is the normal mode through which the student is presented regular print materials and is documented on the student's IEP or 504 . . . plan" (Kentucky Department of Education, 1996, n.p.).

KDE's guidelines for providing a scribe in on-demand parts of the KIRIS Assessment include the following:

A scribe may only be used for the KIRIS assessment when:

- a student has a verified disability in the area of written expression or a physical disability which impedes the motor process of writing;
- a student is motorically able to print or use cursive techniques . . . ; however, the student's written language deficit is so severe that the student cannot translate thoughts into written language even though the student can express those thoughts orally. This is a very rare situation in which such students cannot recognize written words or make sound-symbol associations;
- a student can write, but writes very slowly and the time constraint of the . . . task will inhibit the student's ability to produce the required product. (Kentucky Department of Education, 1996, Attachment G, A6)

A scribe may not be used for the KIRIS Assessment:

- to enhance student products, i.e., the student is able to produce the product, but the product would be better if it were scribed. (Kentucky Department of Education, 1996, Attachment G, A6)

Although careful reading suggests that KDE intended that these criteria be interpreted as very restrictive, the two last criteria might introduce uncertainties into the decision about offering a scribe. At what point does a student's slowness in writing change from merely degrading the quality of the product (indicated to be insufficient grounds for accommodation by the last criterion) to "inhibiting" production of the product (indicated to be sufficient ground for accommodation

by the previous criterion)? The fact that additional time is allowed for most parts of KIRIS might further cloud the decision.

How Many Students With Disabilities Were Tested?

In 1995, Kentucky succeeded in including most students with disabilities in the regular KIRIS assessment, although some had no recorded scores. We estimated that all but 10% to 15% of students with disabilities were listed by the KIRIS assessment program, and all but roughly 15% to 20% of all students with disabilities had KIRIS scores recorded (Koretz, 1997).

We found much the same pattern in 1997, with one potentially important change. While the total number of elementary and middle school students participating in KIRIS dropped slightly over the two-year period, the number identified as disabled increased slightly. In 1995, 10.0% of elementary students and 8.1% of middle school students in the KIRIS database were identified as disabled (Koretz, 1997, Table 2). Two years later, those percentages had increased to 11.2% and 8.8 percent, respectively (Table 1). That is, the percentage of elementary students identified as disabled increased by 12% in the space of only two years, while the percentage of middle school students identified increased by almost 9%.

These changes are modest and may have little significance, but they bear further investigation because they suggest that the KIRIS accountability system may have created unwanted incentives for student classification. By counting students with disabilities in school averages regardless of whether they are tested, KIRIS removes the incentive to exclude such students from testing. The KIRIS approach may also lessen one incentive to over-identify low performing

Table 1

Students With Disabilities Assessed With Regular KIRIS, 1997

	Grade 4	Grade 5	Grade 7	Grade 8	Grade 11
All students tested in KIRIS	48328	47712	50806	50635	41626
Students with disabilities tested in KIRIS	5441	5351	4455	4455	2096
Tested students with disabilities as a percent of all tested students	11.3%	11.2%	8.8%	8.8%	5.0%

students as disabled, since identification does not eliminate a student's influence on scores. However, identifying students as disabled does still confer one advantage to educators trying to maximize scores: It permits them to provide students with assessment accommodations, provided those accommodations are also in the student's IEP. Thus the small changes noted here raise the question of whether educators are responding to this incentive to classify additional students as disabled in order to raise scores.

The use of open-response items appears not to have been a barrier to the inclusion of students with disabilities. In all grades, somewhat more students with disabilities obtained scores on the open-response (OR) portion of the assessment than on the multiple-choice (MC) section. This was true of nondisabled students as well, although less so. Therefore, the percentage of students with scores identified as disabled was slightly higher for the open-response portion than for the multiple-choice portion (Table 2). The KIRIS data do not indicate, however, what the effects of using open-response items would have been if accommodations had been used more sparingly.

Nearly half of all students with disabilities tested with the regular KIRIS assessment were classified as having specific learning disabilities as their primary disability (Table 3). Somewhat more than one fourth were classified as having mild mental retardation. These are the only groups large enough for many of the analyses described in this report. Communication and speech disorders were

Table 2

Students With Disabilities With KIRIS Scores, 1997

	Grade 4	Grade 5	Grade 7	Grade 8	Grade 11
Students with disabilities with MC scores	5160	5194	4329	4144	1965
Students with disabilities with MC scores as a percent of all students with MC scores	10.9%	11.0%	8.7%	8.3%	4.8%
Students with disabilities with OR scores	5440	5351	4455	4454	2096
Students with disabilities with OR scores as a percent of all students with OR scores	11.3%	11.2%	8.8%	8.8%	5.0%

Note. MC = multiple choice; OR = open response.

Table 3

Percentages of All Tested Students With Disabilities in Each Primary Disability Category, 1997

	Grade 4	Grade 5	Grade 7	Grade 8	Grade 11
Autism	0.4	0.3	0.2	0.3	0.1
Deaf/Blind	0.0	0.0	0.0	0.1	0.3
Multiple Disabilities	2	1	1	2	2
Emotional/Behavioral Disabilities	6	8	12	11	7
Mild Mental Disabilities	26	27	29	29	29
Physical Disabilities/Orthopedically Impaired	1	1	0.4	1	1
Other Health Impaired Disabilities	6	6	4	3	2
Traumatic Brain Injury	0.3	0.3	0.2	0.3	1
Hearing Impaired	0.9	1.0	0.9	1.2	2
Visual Disabilities	1	1	1	1	1
Communications/Speech-Language Disabilities	12	6	2	1	0.3
Functional Mental Disabilities	3	2	2	5	5
Specific Learning Disabilities	42	47	47	46	50

Note. Includes only students with scores on the regular KIRIS assessment. Percentages may not sum to 100 because of rounding.

relatively common in the fourth grade but became infrequent as students progressed through the grades. Students with emotional and behavioral disabilities constituted the third largest group in all grades but the fourth grade. It is important to note that the physical disabilities that often figure prominently in discussions of assessment accommodations, such as visual, hearing, and physical/orthopedic disabilities, were very rare in Kentucky. In most grades, these groups each constituted one percent or less of the students with disabilities assessed with KIRIS. These percentages are similar to those found in the earlier study (Koretz, 1997, Table 4). While the reported prevalence rates of specific disabilities vary greatly from state to state, the high prevalence of specific learning disabilities and the very low prevalence of conditions such as visual and hearing disabilities are commonly found nationwide (see U.S. Department of Education, 1996).

How Were Accommodations Used?

In our earlier study, we showed that in 1995, most students with disabilities were given at least one assessment accommodation, and many were given more than one. The use of accommodations was particularly intensive in Grade 4 and declined somewhat in the higher grades. Particular accommodations that one might expect to be used relatively infrequently, such as paraphrasing and taking dictation, were provided to large numbers of students with disabilities, despite the guidelines provided by the Kentucky Department of Education.

Unfortunately, the 1997 KIRIS data indicate only whether a student was provided a given accommodation at some point in the assessment and not whether it was provided specifically for the open-response component, the multiple-choice component, or both. Thus, the data do not show whether accommodations were used differently for the two formats, but they do indicate whether the reincorporation of multiple-choice items altered the frequency with which students were given accommodations for at least part of the assessment.

Neither the additional two years of experience nor the reintroduction of multiple-choice items substantially changed the uses of accommodations noted in 1995. Therefore, we provide here only a brief summary of the use of accommodations as a context for understanding the performance data reported below. Readers interested in more detail about the use of accommodations are referred to the earlier report (Koretz, 1997).

In 1997, fewer than one fifth of the elementary school students with disabilities were assessed without accommodations, while about two thirds were assessed with two or more (Table 4). The percentage assessed without accommodations rose, and the percent assessed with multiple accommodations

Table 4

Percentage of Students With Disabilities Receiving Assessment Accommodations, by Grade, 1997

	Grade 4	Grade 5	Grade 7	Grade 8	Grade 11
No accommodations	19	16	26	31	36
One accommodation	14	19	27	28	26
Multiple accommodations	67	65	47	41	38

fell, as students moved through the secondary grades. These percentages are similar to those recorded two years earlier.

The accommodations most commonly used in 1997 were oral presentation, paraphrasing, and taking dictation of the student's answers (Table 5). Note that these percentages reflect simple counts: All students who received a given accommodation were included in the count for that accommodation, regardless of whether they also received additional accommodations. For example, the 72% of fourth-grade students shown as receiving oral presentation includes both those who received only oral presentation and those who received oral presentation in combination with one or more other accommodations. The use of oral presentation declined modestly with increasing grade level, whereas the use of dictation declined dramatically. Despite the state's guidelines, paraphrasing was provided to roughly half of all students with disabilities in all grades. All of these percentages are similar to those found in 1995. Particularly important for interpreting the performance patterns described below is the finding that the use of dictation increased only slightly from 1995 to 1997 (compare Koretz, 1995, Table 6). For example, 50% of fourth-grade students with disabilities received dictation in 1995, compared to 55% of fourth graders and 49% of fifth graders in 1997.

It is also useful to look at the use of specific, mutually exclusive categories of accommodations, taking the use of multiple accommodations into account. For

Table 5

Percentage of Students With Disabilities Receiving Assessment Accommodations, by Grade, 1997 (Based on Simple Counts)

Accommodation	Grade 4	Grade 5	Grade 7	Grade 8	Grade 11
None	19	16	26	31	36
Oral presentation	72	73	60	54	42
Paraphrasing	48	49	48	46	49
Dictation	55	49	21	15	7
Cueing	10	9	7	8	7
Technological aid	34	4	4	4	4
Interpreter	1	2	1	1	2
Other	9	10	6	7	7

Note. Individual students may receive multiple accommodations.

example, students who received both oral presentation and dictation would be counted as one group and would not be included with the students who received only oral presentation.

Although a large number of these mutually exclusive accommodations were used, only a few of them were provided to sizable numbers of students: oral presentation alone, paraphrasing alone, and several combinations of oral presentation, paraphrasing, and dictation (Table 6). All of these percentages are quite similar to those found in 1995 (Koretz, 1997, Table 8), and the changes are too small to account for the sizable changes in performance described below.

How Do Students With Disabilities Perform on KIRIS?

The following sections provide several different views of the performance of students with disabilities on the KIRIS assessment. The first section provides summary statistics for all students with disabilities, separately for the open-response and multiple-choice components of the assessment but without regard to primary disabilities or the use of accommodations. Subsequent sections examine the performance of specific disability groups and accommodations.

All scores were standardized to make them comparable across grades, subjects, and format (open-response vs. multiple choice). KDE scaled the assessment separately for the two formats, so each student received in each subject one performance estimate for the multiple-choice component and a second estimate for the open-response component. Both estimates were IRT-

Table 6

Percentage of Students With Disabilities Receiving Assessment Accommodations, by Grade (Based on Mutually Exclusive Categories)

	Grade 4	Grade 5	Grade 7	Grade 8	Grade 11
None	19	16	26	31	36
Oral presentation only	8	11	16	16	9
Paraphrasing only	2	4	7	9	14
Oral presentation and dictation	19	17	5	4	1
Oral presentation and paraphrasing	8	11	20	17	19
Oral presentation, paraphrasing, and dictation	22	19	9	6	3
Other multiple accommodations	18	19	13	14	14

based theta scores. We standardized all of the distributions of theta scores to have a mean of 0 and a standard deviation of 1 in the population of students without disabilities. Thus, in normative terms, any given score or group difference reported here has the same meaning regardless of grade, subject, or format. For example, in every instance, a mean difference of .5 between students with and without disabilities would indicate that the mean student with disabilities would rank at about the 31st percentile among students without disabilities in that subject, grade, and format.

All Students With Disabilities

On average, students with disabilities scored well below students without disabilities in every case, but there were important variations across grade, subject, and format. (See Table 7. Note that in all tables, MC denotes multiple choice, and OR denotes open response.) In most cases, the gap between the groups was .9 standard deviation or larger, but it increased across the grades. The smallest difference was .4 standard deviation in fourth-grade open-response

Table 7

KIRIS Means (and Standard Deviations) for All Students With Disabilities, by Subject, Format, and Grade, 1997

	Reading	Math	Science	Social Studies
Grade 4 (*N*=5160)				
MC	-0.5 (0.9)		-0.7 (1.0)	
OR	-0.5 (1.1)		-0.4 (1.1)	
Grade 5 (*N*=5194)				
MC		-0.9 (1.0)		-1.0 (1.0)
OR		-0.7 (1.0)		-0.7 (1.1)
Grade 7 (*N*=4329)				
MC	-1.0 (0.9)		-0.9 (1.0)	
OR	-1.1 (1.0)		-0.9 (1.0)	
Grade 8 (*N*=4143)				
MC		-1.1 (0.8)		-1.0 (0.9)
OR		-1.1 (0.9)		-1.0 (0.9)
Grade 11 (*N*=1965)				
MC	-1.4 (0.9)	-1.1 (0.7)	-1.0 (0.8)	-1.0 (0.9)
OR	-1.4 (1.0)	-1.0 (0.7)	-1.3 (0.9)	-1.3 (0.9)

Note. MC = multiple choice; OR = open response. Only students with scores on both MC and OR tests are included.

science, which would place the mean student with disabilities at roughly the 34th percentile among students without disabilities. The largest disparity was in Grade 11 reading; in that case, the difference of 1.4 standard deviations (approximately the same in both formats) would place the mean student with disabilities at around the 8th percentile among students without disabilities.

The relationships between format and performance were complex. In the two elementary grades, students with disabilities tended to score more poorly (relative to students without disabilities) on the multiple-choice component than on the open-response component. This difference appeared in mathematics, science, and social studies. The exception was fourth-grade reading, in which students with disabilities obtained roughly the same average scores on the two components. This format difference had disappeared by the middle school grades. In Grades 7 and 8, students with disabilities scored about the same (again, relative to students without disabilities) on both formats, regardless of subject. The format difference apparent in the fourth grade was partially reversed in Grade 11. Eleventh-grade students with disabilities scored the same on both formats in reading and trivially lower on the multiple-choice component in mathematics, but they scored more poorly on the open-response component in both science and social studies.

The data do not indicate why the gap between students with and without disabilities grew and the effects of format changed as students progressed through the grades, but there are at least three possible explanations. First, the population of students with disabilities differs from grade to grade. (Note the declining percentages of tested students with disabilities in Table 2.) Second, some students with disabilities may fall progressively farther behind as they progress through the grades, and their growing performance deficit may not appear equally in tasks using different formats. Third, differences in the use of accommodations across the grades (Table 5 above) may also contribute to these patterns—in particular, to the larger mean differences in the higher grades.

These results indicate a considerable worsening of the performance of elementary school students with disabilities over the two years since the earlier study. In 1997, the mean differences on the open-response component in the fourth and fifth grades ranged from .4 standard deviation in science to .7 standard deviation in both mathematics and social studies (Table 7 above). In contrast, the corresponding differences in the fourth grade two years earlier

ranged from .1 standard deviation in science to .4 standard deviation in mathematics (Koretz, 1997, Table 9). The performance gap in the middle school and high school grades, on the other hand, did not change markedly. The causes of this change in the elementary grades are not clear, but it may be related to a change in the performance correlates of accommodations noted below.

Performance of Students With Specific Disabilities

We tabulated performance by subject and format for the four largest disability groups: specific learning disabilities, mild mental retardation, and emotional/behavioral disabilities in all grades, and communication/speech disorders in the fourth and fifth grades. Students with and without accommodations were included in these analyses, but only if they had scores reported for both the multiple-choice and open-response components of the assessment.

In all cases, the average scores of students with disabilities were lower than those of students without disabilities, but the size of the performance gap varied dramatically: from a low of .1 standard deviation for 4th-grade students with learning disabilities on the open-response science assessment to a high of 1.7 standard deviations in reading for 11th-grade students with mild mental retardation. One would expect students with mental retardation to show larger performance gaps, but the size of the gap appears to be reflect a complex interaction of disability, grade, and format.

In the elementary grades, the gap between students with learning disabilities and students without disabilities ranged from .1 to .7 standard deviation, depending on subject and format (Table 8). In reading, the average scores of students with learning disabilities were the same for both formats, but in all other subjects, their average score on the multiple-choice component was .3 standard deviation lower than their performance on the open-response portion. Note that these differences are all normative. That is, they indicate only that the gap between students with learning disabilities and those with no disability were larger on the multiple-choice portion of the assessment.

The scores of elementary students with mild mental retardation were predictably much lower than those of students with learning disabilities. Their average scores ranged from .8 to 1.5 standard deviations lower than those of students without disabilities (Table 8). Again, scores were lower on the multiple-

Table 8

Mean Scores by Disability and Subject, Grades 4 and 5

	No. tested	Reading	Math	Science	Social studies
Specific learning disability					
Grade 4	2278				
MC		-0.3		-0.4	
OR		-0.3		-0.1	
Grade 5	2486				
MC			-0.7		-0.7
OR			-0.4		-0.4
Mild mental retardation					
Grade 4	1345				
MC		-1.0		-1.4	
OR		-0.9		-0.8	
Grade 5	1399				
MC			-1.5		-1.4
OR			-1.2		-1.0
Emotional/behavioral					
Grade 4	322				
MC		-0.5		-0.7	
OR		-0.7		-0.6	
Grade 5	423				
MC			-0.9		-1.0
OR			-0.8		-0.8
Communication/speech					
Grade 4	643				
MC		-0.6		-0.6	
OR		-0.5		-0.5	
Grade 5	330				
MC			-0.6		-0.8
OR			-0.6		-0.5

Note. Scores are scaled to a mean of 0 and a standard deviation of 1 in the population of students without disabilities. Only students with scores on both MC (multiple choice) and OR (open response) tests are included.

choice format than on the open-response portion; the format difference ranged from .1 standard deviation in reading to .6 standard deviation in science. In the elementary grades, scores for the other two disability groups, emotional/behavioral and communication/speech disabilities, were higher than those of

students with mental retardation but similar to or larger than those of students with learning disabilities (Table 8 above).

Several patterns in the scores of elementary students with disabilities bear mention. First, even though most students with learning disabilities are identified as having a reading disability, these students scored appreciably *higher* than students in all of the three other disability groups in reading, regardless of format. Second, while students with emotional/behavioral and communication/speech disabilities tended also to have lower scores on the open-response component, this format difference was generally smaller for these groups than for others, and there was one exception (in fourth-grade reading). Third, a comparison with scores from two years earlier (compare Koretz, 1997, Table 10) shows that the decline in performance over the two years, while apparent in all four groups, was generally more modest among students with emotional/behavioral and communication/speech disabilities.

As noted above, the average performance of students with disabilities—relative to students without disabilities—dropped between the elementary and middle school grades. This decline was apparent for both specific learning disabilities and emotional/behavioral disabilities but was less consistent for students with mild mental retardation (Table 9).[2] The format differences that appeared clearly in the elementary grades are not generally present in the middle school grades. In every subject, the average scores of middle school students with specific learning disabilities or mild mental retardation on the two components (multiple choice and open response) were within .1 standard deviation. Middle school students with emotional/behavioral disabilities had *lower* scores on the open-response component, by as much as .3 standard deviation (in reading and science).

The relative performance of students with learning disabilities dropped further from the middle school to the high school level (Table 10). For the other two largest disability groups, the change between these grades was inconsistent, although there were some instances (such as open-response science for students with either mild mental retardation or emotional/behavioral disabilities) in

---

[2] Students with communication or speech disorders are not included in tables for the middle school or high school levels because of their small numbers in those grades.

Table 9

Mean Scores by Disability and Subject, Grades 7 and 8

	No. tested	Reading	Math	Science	Social studies
Specific learning disability					
Grade 7	2093				
MC		-0.9		-0.7	
OR		-0.9		-0.8	
Grade 8	2044				
MC			-1.0		-0.9
OR			-0.9		-0.8
Mild mental retardation					
Grade 7	1286				
MC		-1.3		-1.3	
OR		-1.3		-1.2	
Grade 8					
MC	1259		-1.4		-1.4
OR			-1.4		-1.3
Emotional/behavioral					
Grade 7	534				
MC		-1.0		-0.9	
OR		-1.3		-1.2	
Grade 8	457				
MC			-1.1		-1.1
OR			-1.3		-1.2

Note. Scores are scaled to a mean of 0 and a standard deviation of 1 in the population of students without disabilities. Only students with scores on both MC (multiple choice) and OR (open response) tests are included.

which average scores dropped appreciably. As noted earlier, among all students with disabilities, the elementary school format difference, in which students with disabilities scored more poorly on the multiple-choice component, was reversed in the 11th grade. The lower performance of high school students on the open-response component was apparent in all three of the largest disability groups, but it was largest and most consistent among students with emotional/behavioral disabilities.

Table 10

Mean Scores by Disability and Subject, Grade 11

	No. tested	Reading	Math	Science	Social studies
Specific learning disability	1037				
MC		-1.3	-1.0	-1.0	-1.0
OR		-1.3	-1.0	-1.2	-1.2
Mild mental retardation	577				
MC		-1.7	-1.3	-1.3	-1.3
OR		-1.7	-1.3	-1.6	-1.5
Emotional/behavioral	142				
MC		-1.1	-0.9	-0.8	-0.9
OR		-1.5	-1.1	-1.5	-1.4

Note. Scores are scaled to a mean of 0 and a standard deviation of 1 in the population of students without disabilities. Only students with scores on both MC (multiple choice) and OR (open response) tests are included.

Performance of Disabled Students With and Without Accommodations

The drop in performance of elementary students with disabilities from 1995 to 1997 reflects changes in the performance of students who received accommodations.

The performance of students without accommodations on the open-response portion of the assessment in 1997 was generally similar to that in 1995 (Table 11; compare Koretz, 1997, Table 12). For example, in 1997, elementary students with disabilities who were tested with no accommodations scored 0.6 to 0.8 standard deviation below their nondisabled peers on the open-response portions of the assessment, depending on the subject (Table 11). Two years earlier, the corresponding gaps for fourth-grade students with disabilities ranged from 0.6 to 0.7 standard deviation.

The performance of elementary school students with disabilities who received accommodations, however, dropped markedly. In 1997, the means of elementary disabled students with accommodations ranged from 0.4 to 0.7 standard deviation below the mean for nondisabled students (Table 11). In contrast, the means for fourth-grade disabled students with accommodations in 1995 ranged from 0.1 *above* the mean for nondisabled students to 0.3 standard

Table 11

Number and Mean Scores for All Students with Disabilities, by Subject and Grade, With and Without Accommodations

	No. tested	Reading		Math		Science		Social studies	
		MC	OR	MC	OR	MC	OR	MC	OR
Grade 4									
No accommodations	785	-0.7	-0.7			-0.6	-0.6		
Any accommodations	4375	-0.5	-0.5			-0.7	-0.4		
Grade 5									
No accommodations	704			-0.8	-0.8			-0.8	-0.7
Any accommodations	4490			-1.0	-0.7			-1.0	-0.6
Grade 7									
No accommodations	1031	-0.9	-1.0			-0.7	-1.0		
Any accommodations	3298	-1.0	-1.1			-0.9	-0.9		
Grade 8									
No accommodations	1087			-1.0	-1.1			-0.9	-1.0
Any accommodations	3056			-1.1	-1.0			-1.1	-1.0
Grade 11									
No accommodations	627	-1.2	-1.3	-1.0	-1.1	-0.8	-1.3	-0.9	-1.2
Any accommodations	1338	-1.5	-1.4	-1.1	-1.0	-1.1	-1.3	-1.1	-1.3

Note. Scores are scaled to a mean of 0 and a standard deviation of 1 in the population of students without disabilities. Only students with scores on both MC (multiple choice) and OR (open response) tests are included.

deviation below (Koretz, 1997, Table 12). This decline did not appear in the scores of older disabled students tested with accommodations.

It was noted earlier that elementary students with disabilities tended to score lower on multiple-choice items than on open-response items in all subjects other than reading. This format difference is almost entirely attributable to students with accommodations (Table 11). In the 11th grade, students with disabilities tended to score lower on the open-response portion. That format difference, however, is largely attributable to students *without* accommodations.

Performance for Mutually Exclusive Categories of Accommodations

The performance of disabled students when classified by the accommodations they received was one of the primary reasons why we questioned the validity of some scores in our earlier report (Koretz, 1997).

Specifically, learning-disabled and mildly retarded students receiving certain specific combinations of accommodations, all of which included dictation, received implausibly high scores: nearly average in the case of students with mild retardation, and well above average in the case of students with learning disabilities.

Figure 1 shows the 1997 mean scores of elementary school learning-disabled students who received either no accommodations or one of the five most common mutually exclusive categories of accommodations: oral presentation alone, paraphrasing alone, paraphrasing and oral presentation together, oral presentation and dictation together, or paraphrasing, oral presentation, and dictation together. The last two groups—those that received dictation in combination with other accommodations—are the two that generally scored well above average in 1995 (Koretz, 1997, Figure 2).

In contrast, the average scores of these groups of learning-disabled students on the open-response components were near or below the means for nondisabled students in 1997 (Figure 1). Specifically, the means for disabled students with these accommodations ranged from slightly above the mean in fourth-grade science to 0.5 standard deviation below the mean in mathematics (for the group that received only oral presentation and dictation). These mean scores seem more plausible than the higher means found in 1995.

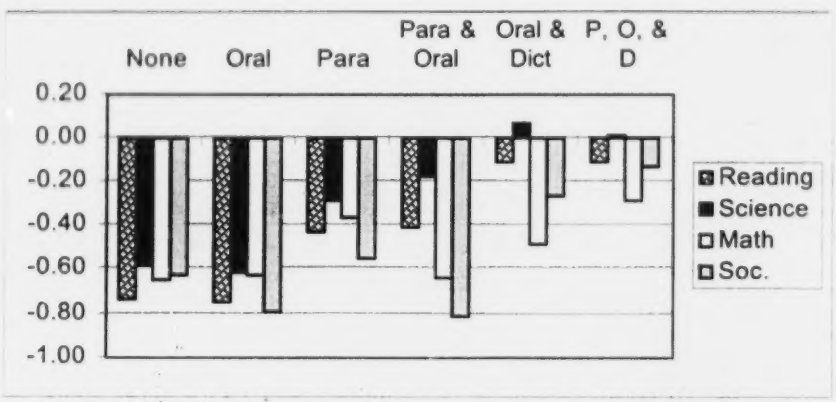


Figure 1. Elementary learning-disabled students, open-response means by accommodations.

The performance correlates of accommodations for learning-disabled students are quite different in the case of the multiple-choice components (Figure 2). The performance of students without any accommodations was quite similar for both formats, roughly 0.6 to 0.8 standard deviation below the mean, depending on the subject, in both cases (compare Figures 1 and 2). The scores of students receiving combinations of accommodations that include dictation were much lower on the multiple-choice components than on the open-response components, particularly in mathematics and social studies. This is not surprising: While oral presentation and paraphrasing could be germane to both components, dictation was presumably irrelevant for the multiple-choice component. The relationship between the other combinations of accommodations and performance are inconsistent. Learning-disabled students receiving paraphrasing scored higher on the open-response component, whereas in two subjects (reading and science), those that received oral presentation only scored higher on the multiple-choice component.

The 1997 scores of elementary school mildly retarded students receiving dictation also dropped markedly from 1995. In 1997, mildly retarded elementary school students who received these two specific combinations of accommodations—dictation in combination with other accommodations—scored well below the mean for nondisabled students (Figure 3). Their average scores ranged from 0.6 to 1.2 standard deviations below the mean for nondisabled students, and all but one of the eight means were at least 0.8 standard deviation below the mean. In contrast, the average scores of these same groups in 1995

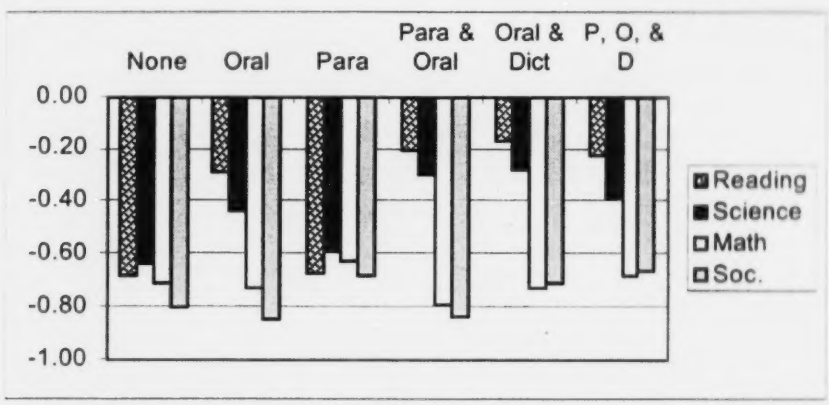


Figure 2. Elementary learning-disabled students, multiple-choice means by accommodations.

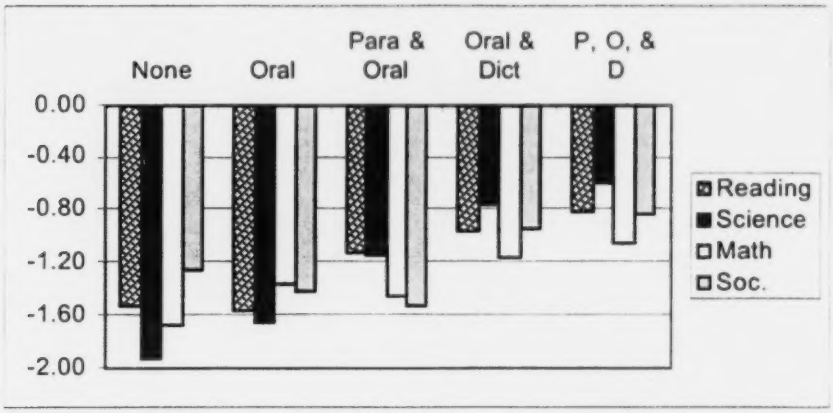


Figure 3. Elementary mildly mentally handicapped students, open-response means by accommodations.

ranged from 0.7 standard deviation below the mean for nondisabled students to 0.1 standard deviation *above* the mean, and only two of the eight scores were 0.4 standard deviations or more below the mean. Given that students with mild retardation have generalized cognitive deficits, the lower scores obtained in 1997 again appear more plausible.

The performance of students with disabilities in the secondary grades varied less as a function of the accommodations they were given. For example, in the middle school grades, learning-disabled students receiving dictation in combination with other combinations scored substantially higher than other learning-disabled students on the open-response science component but not on the open-response mathematics component (Figure 4). Moreover, in all cases, the means for learning-disabled students were well below the average for nondisabled students. In the 11th grade, there were too few learning-disabled students receiving dictation in combination with other accommodations to permit this type of analysis.

The performance of middle school students with learning disabilities on the multiple-choice components of the examination was similar to that on the open-response components, except that the means for the two groups receiving dictation differed even less from the means of other groups. In science, however, those two groups did have appreciably higher means, a finding that may reflect confounding with differences among these groups of students rather than effects of accommodations. This possibility is explored further in the following section.

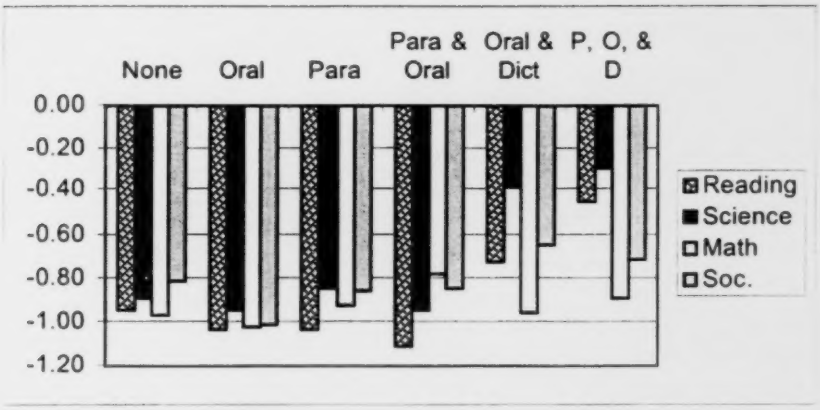


Figure 4. Learning-disabled students, middle school open-response means by accommodations.

Performance Correlates of Individual Accommodations

Most students with disabilities were given two or more accommodations, and the small groups of students receiving any single accommodation alone do not provide a reasonable basis for inferring their effects. Therefore, as in the earlier study, we used multiple regression (simple ordinary least squares) to disentangle the independent associations between individual accommodations and KIRIS scores. These analyses were carried out for the two largest groups, students with learning disabilities and mild mental retardation, separately for the open-response and multiple-choice components of the assessment. In both years, the results were quite similar for learning-disabled and mildly retarded students. Therefore, for simplicity, we focus primarily on the results for the larger, learning-disabled group.

As we warned in our earlier study, we cannot be certain that these regression estimates represent the actual effects of accommodations because the data are not experimental. These estimates may be influenced by other differences among the groups of students assigned various accommodations. It seems reasonable to assume that in general, students given the most substantial accommodations are those with the most severe disabilities. If true, that would tend to mask the effects of accommodations. That is, the positive effects of the accommodations on scores would be partially offset by the negative effects of disabilities. This in turn suggests that the regression estimates are likely to

understate the effects of accommodations. This need not always be so, however, and we will discuss one instance in which it appears not to be.

The results of the regression analysis of 1997 data appear to explain much of the decline since 1995 in the mean performance of elementary school students with disabilities. They also suggest limitations of the current, non-experimental approach to analyzing the correlates of accommodations.

Our earlier analysis of the 1995 KIRIS data showed that the independent, positive association of dictation with performance was much larger than that of any other single accommodation. In the case of learning-disabled students, for example, the mean and median estimates for dictation were both about 0.7 standard deviation, and most were greater than 0.6 standard deviation. (See Koretz, 1997, Figure 5 and Appendix A.) The estimates for cueing were second-largest but were much smaller; the largest estimates were somewhat over 0.3 standard deviation. The estimates for paraphrasing exceeded 0.2 standard deviation on only two of 12 cases, and the estimates for oral presentation were always well under 0.2 standard deviation.

Two years later, in 1997, the estimated association between dictation and performance for learning-disabled students, while still large—and larger than those of the other common accommodations—had shrunk considerably. Because roughly half of the elementary school students with disabilities received dictation, the apparent decline in the impact of dictation could help explain the sizeable drop in their mean scores on KIRIS. The median estimated was about 0.5 standard deviation, and the mean was about 0.4 standard deviation. In 1997, most of the estimated effects clustered between 0.4 and 0.6 standard deviation, and three were much smaller. (See Figure 5, in which each symbol represents a single coefficient for dictation, and the size of the coefficient is plotted on the x-axis. Note that there are twelve cases for each year: three grades by four subjects.) The rank ordering of the coefficients also changed between 1995 and 1997. In 1997, for example, the three unusually small effects of dictation were in 8th-grade mathematics, 8th-grade social studies, and elementary (5th-grade) mathematics (listed from smallest to largest). The only one of these that showed a relatively small association in 1995 was elementary (4th-grade) mathematics. Other estimates remained more similar; for example, 11th-grade social studies and science had estimates that were among the largest in both years.

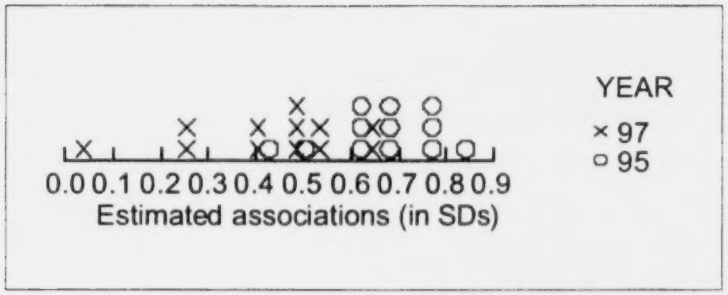


Figure 5. Estimated associations between dictation and open-response scores, learning-disabled students, 1995 and 1997.

Dictation generally had a larger association with the open-response scores of learning-disabled students than with their multiple-choice scores (Figure 6). One might expect only trivial effects of dictation on multiple-choice scores, however, and thus the fact that some of the estimates for the multiple-choice components are in the range of 0.3 standard deviation may suggest errors in the regression model. These errors might be of several types. For example, it may be that students receiving dictation have other characteristics that cause them to achieve higher scores—although this would contradict our assumption that students who receive accommodations will in general be lower performing than those who do not. Alternatively, some students who receive dictation may be receiving additional assistance not captured by the other accommodations variables in the data.

In 1997, as in 1995, cueing—which was provided to few students (Table 5)—showed the second-largest association with the open-response scores of learning-disabled students. This association ranged from zero to 0.36 standard

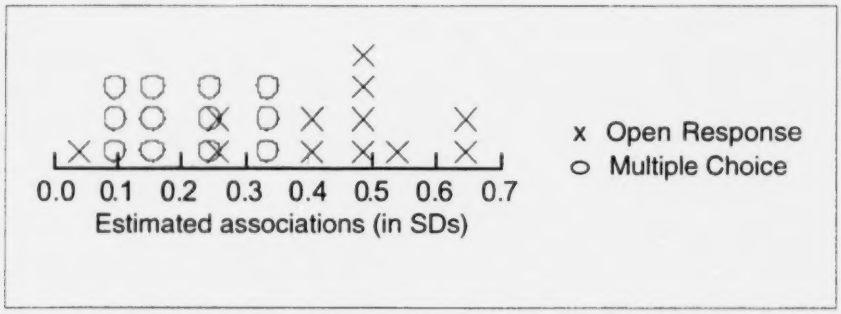


Figure 6. Estimated associations between dictation and scores, learning-disabled students, 1997, by format.

deviation, but most were clustered around 0.3 standard deviation, and the mean and median were both approximately 0.3 standard deviation (Figure 7). With the exception of one case, cueing showed only very weak associations with multiple-choice scores. This raises important questions about the nature of the cueing assistance students were provided. It suggests that the cueing pertained primarily to responses rather than to analysis of items and prompts.

Oral presentation showed an appreciable positive association with the performance of learning-disabled students in only one case: fourth-grade reading, where the association was a bit over 0.3 standard deviation (the highest point in Figure 8). In several instances, however, the estimated associations were modestly negative—that is, providing oral presentation was associated with scores that were lower by roughly 0.2 standard deviation. In some cases, these were consistent for a given group of students. In particular, the fifth-grade sample provided the most strongly negative estimates for the open-response components and two of the three most negative estimates for the multiple-choice components. It is possible that oral presentation actually hinders enough students that these estimates are correct; for example, it may impede the performance of some students by slowing them down or by distracting them from a written presentation that offers the opportunity for review. On the other hand, it is also possible that the true effects of oral presentation are not negative in these cases, and that the negative estimates reflect an inadequacy of the regression models. For example, it may be that the impact of oral presentation is too small to overcome the performance deficits of some students who are offered that accommodation.

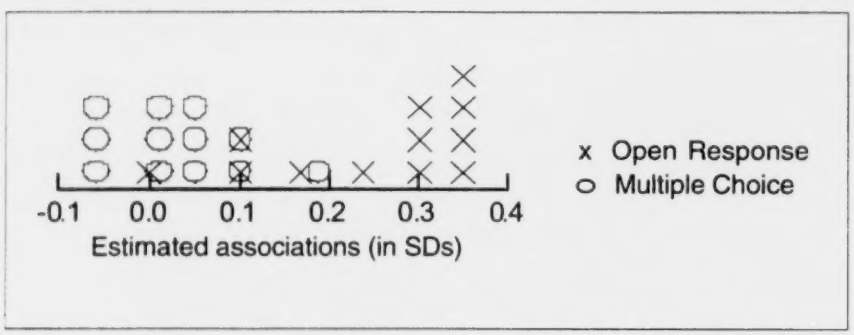


Figure 7. Estimated associations between cueing and scores, learning-disabled students, 1997, by format.

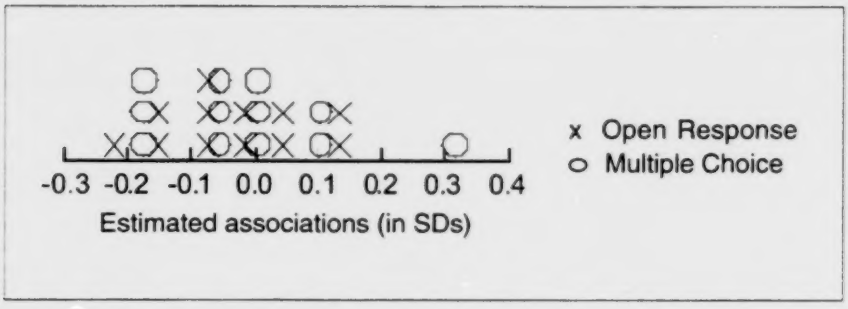


Figure 8. Estimated associations between oral presentation and scores, learning-disabled students, 1997, by format.

In 1997, paraphrasing showed largely trivial associations, both positive and negative, with the open-response scores of learning-disabled students (Figure 9). The largest associations were 0.15 standard deviation, in eighth-grade mathematics and fourth-grade reading. This stands in some contrast to 1995, when all of the estimates were positive and the largest were over 0.2 standard deviation (Koretz, 1997, Figure 5). The associations between paraphrasing and multiple-response scores were all near-zero or negative. In this case, the negative estimates are small, but this again suggests some weaknesses in the models.

The estimates for mildly retarded students showed largely similar patterns. The estimated associations between cueing and scores were on average larger for mildly retarded students than for learning-disabled students, although in this group as well, the estimates tended to be considerably larger for the open-response component than for the multiple-choice components. The estimates for dictation, however, were smaller for mildly retarded students, particularly in the case of the open-response components. Nonetheless, the estimates for dictation

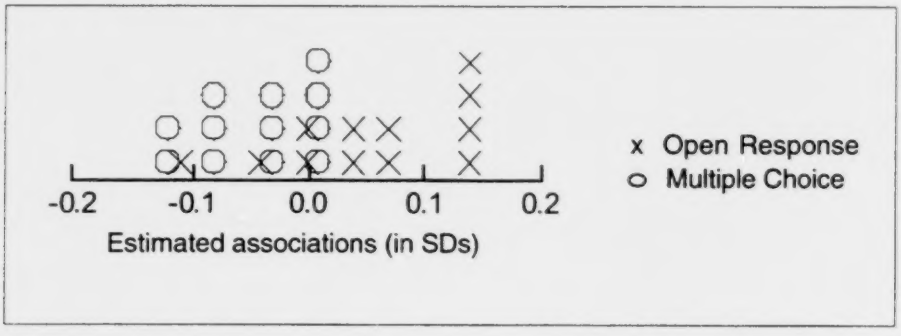


Figure 9. Estimated associations between paraphrasing and scores, learning-disabled students, 1997, by format.

were the largest for mildly retarded students, averaging roughly 0.4 standard deviation in the case of the open-response components.

Correlations Among Parts

This section presents the correlations among the multiple-choice (MC) and open-response (OR) sections of the tests within each grade, separately for nondisabled students, students with disabilities who received accommodations, and students with disabilities who did not receive accommodations. Of primary interest was whether the relationships among the tests in different subjects and formats varied by group. It would be possible, for example, for science scores to be more heavily dependent upon reading ability in some groups than in others, or for multiple-choice and open-response tests to function more or less similarly for different groups.

All of the parts were moderately correlated with one another in all grades and for all groups. However, for all but two pairs of tests, the correlations among parts for accommodated students with disabilities were lower than those for the other two groups. The exceptions were fourth-grade OR science with OR reading, and fifth-grade OR math with OR social studies, both of which showed slightly higher correlations for accommodated disabled students than for nondisabled students. The differences among groups were trivial in many cases, and substantial in a few. For example, Table 12 presents the correlations among parts for seventh grade, separately by group. Correlations for nondisabled students and disabled students tested without accommodations were similar to each other in most cases, whereas the correlations for accommodated students with disabilities were more than .05 lower than corresponding correlations for both other groups in all but one instance.

The average differences in correlations between nondisabled students and each of the other two groups are presented in Table 13. The magnitudes of the difference between nondisabled students and students with disabilities who received accommodations increased with grade level, probably due in part to the decreasing score variance among accommodated students. This is not the only explanation for the differences across groups, however, because the differences in variances across groups were small, and in some cases the disabled students had a larger variance than the nondisabled students. The lower correlations for accommodated students may also have resulted from differential use of

Table 12

Correlations Among Parts, Grade 7

		OR		MC	
		Reading	Science	Reading	Science
Nondisabled students					
OR	Reading	1			
	Science	.64	1		
MC	Reading	.67	.65	1	
	Science	.50	.64	.68	1
Students with disabilities, no accommodations					
OR	Reading	1			
	Science	.65	1		
MC	Reading	.65	.60	1	
	Science	.51	.63	.69	1
Students with disabilities, accommodations					
OR	Reading	1			
	Science	.63	1		
MC	Reading	.57	.51	1	
	Science	.44	.55	.62	1

Table 13

Differences Between Nondisabled and Disabled Students in Average Correlations Among Parts

	Difference in average correlations	
Grade	Nondisabled minus unaccommodated, disabled	Nondisabled minus accommodated, disabled
4	-.05	.03
5	-.04	.02
7	.01	.08
8	.05	.17
11	.02	.18

accommodations across parts. It is likely, for example, that dictation was used on the OR tests but not on MC, and this might result in a weaker relationship between OR and MC scores for those students who received dictation.

Unfortunately, as mentioned earlier, we do not have data regarding the specific parts on which accommodations were used.

Another difference between accommodated students with disabilities and the other groups was that the former tended to show stronger format effects, particularly in the case of OR scores. The correlations between OR tests in different subjects tended to be larger than the correlations between OR and MC tests in the same subject for accommodated students but not for other students, especially at the lower grades. In 7th grade, for example, shown in Table 12, the correlation between OR tests in different subjects was .63 for accommodated students, whereas the correlations between OR and MC tests in the same subjects were .57 and .55 for this group. The other two groups did not exhibit this pattern. This may reflect a use of accommodations that influenced OR scores in similar ways across subjects. The magnitudes of the differences in this format effect between accommodated students and the other groups were similar in all grades except Grade 11, where the format effect was substantially larger (e.g., differences of .20 or more between same-format/cross-subject and same-subject/cross-format correlations) and was observed across all groups. In all groups, correlations between MC tests in different subjects were high, probably due in part to the greater reliability of the MC tests.

These results provide additional evidence that the tests may function differently for students who did and did not receive accommodations. More detailed data on how accommodations may have been applied differentially across parts would aid in the interpretation of these correlations. The next set of analyses also explores ways in which the tests function differently across disability and accommodation conditions, but at the level of the individual item.

Item-Level Analyses

All of the results presented to this point have focused on total scores on the MC and OR tests. These analyses revealed potentially important differences in student performance across subjects and formats, but a more complete view of the adequacy of the assessment for students with disabilities requires examining how individual items function for students with and without disabilities. In this section we present simple descriptive statistics and item-test correlations for common items. We also describe the results of a set of differential item functioning (DIF) analyses of these items. The DIF studies were conducted to

identify items on which students with and without disabilities who were matched on total test scores performed differently.

We conducted these analyses for all common items in all subjects and grades and for both formats. The MC tests included 16 common items at each grade and subject, and the OR tests included four. All analyses were conducted using three groups: nondisabled students, students with disabilities tested with accommodations, and students with disabilities tested without accommodations.

Item Difficulty

One way to gauge difficulty is by examining the average item score obtained by students who provided responses to the item. For OR items, this is the mean score on the 5-point scale, and on the MC items, it is the *p*-value, that is, the proportion of respondents who answered the item correctly.

The OR common items were more difficult for students with disabilities than for nondisabled students. The mean scores on each common item, averaged over the set of common items in each subject and grade, are presented in Table 14. Nondisabled students scored from 0.3 to 0.4 points higher than disabled students at Grades 4 and 5, and this gap was even larger in the higher grades. The differences between scores obtained by accommodated and unaccommodated students with disabilities were close to zero in many cases and did not exceed 0.2, so the gap between these groups was much smaller than was observed in 1995. This is consistent with findings reported earlier regarding smaller differences in total scores between these two groups. The science and social studies items tended to be more difficult than the reading items for most students, particularly in the higher grades. The 11th-grade math test was especially difficult for both disabled and nondisabled students.

Results were similar for the MC test. The items tended to be somewhat easier for nondisabled students than for students with disabilities, though these differences were modest (Table 15). In only one case, 11th-grade reading, was the difference between nondisabled students and one of the groups of disabled students larger than 0.2. Here again, differences between disabled students with and without accommodations were generally close to zero.

Table 14

Mean Scores on Open-Response Common Items, by Disability Status and Accommodations (Means of Item-Level Means)

	Reading	Math	Science	Social studies
Grade 4				
No disability	2.0		1.8	
Students with disabilities, no accommodations	1.7		1.4	
Students with disabilities, with accommodations	1.7		1.6	
Grade 5				
No disability		1.9		1.6
Students with disabilities, no accommodations		1.4		1.3
Students with disabilities, with accommodations		1.4		1.3
Grade 7				
No disability	2.1		1.5	
Students with disabilities, no accommodations	1.5		0.9	
Students with disabilities, with accommodations	1.4		0.9	
Grade 8				
No disability		2.1		1.5
Students with disabilities, no accommodations		1.2		0.9
Students with disabilities, with accommodations		1.2		0.9
Grade 11				
No disability	2.4	1.0	1.7	1.6
Students with disabilities, no accommodations	1.4	0.4	0.9	0.8
Students with disabilities, with accommodations	1.3	0.4	0.9	0.7

Another indicator of item difficulty is the percentage of students leaving the item blank or, in the case of OR items, obtaining a score of 0, which indicates an answer that is "totally incorrect or irrelevant."[3] These percentages can be averaged across all common items within a format and subject area.

In the case of OR, the percentages of nondisabled students who omitted or scored zero on common items varied markedly among subjects and was large in some instances (Table 16). In almost all cases, the percentages of disabled students who omitted items or scored zero were much higher than the percentages for nondisabled students, exceeding 30% in 6 of 12 cases and reaching over 70% in the case of Grade 11 math. Disabled students with and without accommodations were similar in this respect. Reading was the least difficult subject by this

[3] For the sake of simplicity, we pooled the relatively small number of cases in which students omitted an open-response item with the scores of zero on that item.

Table 15

Mean Proportion Correct on Multiple-Choice Common Items, by Disability Status and Accommodations (Means of Item-Level Proportions)

	Reading	Math	Science	Social studies
Grade 4				
No disability	0.7		0.7	
Students with disabilities, no accommodations	0.6		0.6	
Students with disabilities, with accommodations	0.6		0.6	
Grade 5				
No disability		0.7		0.7
Students with disabilities, no accommodations		0.5		0.6
Students with disabilities, with accommodations		0.5		0.6
Grade 7				
No disability	0.7		0.6	
Students with disabilities, no accommodations	0.5		0.5	
Students with disabilities, with accommodations	0.5		0.5	
Grade 8				
No disability		0.6		0.6
Students with disabilities, no accommodations		0.4		0.5
Students with disabilities, with accommodations		0.4		0.4
Grade 11				
No disability	0.7	0.5	0.6	0.5
Students with disabilities, no accommodations	0.5	0.3	0.4	0.3
Students with disabilities, with accommodations	0.4	0.3	0.4	0.3

measure for all groups. Compared with the 1995 data, the percents scoring zero in math dropped substantially for elementary and middle school students, whereas the percent scoring zero in reading increased slightly for middle school students.

These averages mask some potentially important variation among items within a grade and subject. For example, on one item on the 4th-grade science test, 11% of the nondisabled students received zeros or omitted the item (compared with 5% across all the items), as did 23% of students with disabilities tested without accommodations (compared with 15% across items). This item was the last one administered in a testing session, so it is possible that many students did not have enough time to produce adequate responses even though proctors are permitted to offer additional time to any student who needs it. This item also required more writing than the other items on this test and was

Table 16

Mean Percent of Students Receiving a Score of Zero (or Omit) on Common Items, by Disability Status and Accommodations

	Reading	Math	Science	Social studies
Grade 4				
No disability	3		5	
Students with disabilities, no accommodations	9		15	
Students with disabilities, with accommodations	7		9	
Grade 5				
No disability		19		7
Students with disabilities, no accommodations		34		19
Students with disabilities, with accommodations		30		16
Grade 7				
No disability	5		14	
Students with disabilities, no accommodations	17		37	
Students with disabilities, with accommodations	16		35	
Grade 8				
No disability		9		13
Students with disabilities, no accommodations		29		37
Students with disabilities, with accommodations		27		35
Grade 11				
No disability	4	38	9	16
Students with disabilities, no accommodations	23	75	35	47
Students with disabilities, with accommodations	19	72	33	50

somewhat less structured, requiring students to produce a list of steps in an experimental design. This requirement may have increased the item's difficulty and discouraged many students from providing even partial responses.

In contrast to the results for OR items, the percentages of students leaving MC items blank were generally small. However, students with disabilities but no accommodations had higher omit rates than the other two groups. For example, in Grade 4, nondisabled students and disabled students with accommodations had, on average, a 0.4% rate of omitting common MC reading items, whereas an average of 1.3% of disabled students without accommodations omitted items. Similar patterns were observed across all grades. Unfortunately we have no way of knowing whether different accommodations were used on the MC and OR portions of the test. However, these results suggest that accommodations may

have affected the responses of disabled students even though there is little evidence of effects on item-level or total scores.

Item-to-Total-Score Correlations

The correlation between an individual item and total test score is often used as a measure of the degree to which an item discriminates between high- and low-performing examinees. Comparing item-test correlations for students with and without disabilities can reveal whether some items are more or less discriminating for students with disabilities than for the remainder of the examinee population, and thereby can provide information regarding the quality of the measurement provided by the items for each group. As in the earlier report (Koretz, 1997), we calculated point-polyserial correlations between scores on the common OR items and theta scores on the OR test for the relevant grade and subject. For the MC items, we calculated point-biserial correlations between item score and MC theta score. Correlations were examined separately in three groups: nondisabled students, disabled students who received accommodations, and disabled students who did not receive accommodations.

Among the 48 OR items examined, correlations between the item and theta scores rarely differed across groups by more than 0.05, and in no case did the difference exceed 0.1. Table 17 shows the correlations for 4th grade reading and science items. Surprisingly, where small differences were observed, the smallest correlation was obtained for the nondisabled group. This pattern did not hold across all grade levels. In some cases, one or both of the groups containing students with disabilities showed smaller correlations than the nondisabled group. However, the magnitudes of correlations were remarkably consistent across groups in all grade levels and subjects, suggesting that the items are approximately equally discriminating for all students.

Table 17

Correlations Between Open-Response Common Item Performance and Theta Scores, Grade 4

	Reading item				Science item			
	1	2	3	4	1	2	3	4
Nondisabled	.66	.70	.77	.74	.70	.67	.59	.69
SWD, unaccommodated	.73	.76	.80	.73	.74	.72	.65	.74
SWD, accommodated	.72	.76	.80	.75	.74	.71	.64	.74

Note. SWD = students with disabilities.

In contrast, the point-biserial correlations calculated for the MC items showed substantial variation across groups of students. In almost all grade/subject combinations, the differences among the three groups ranged from close to zero to between .12 and .20, and a few differences were even larger. Table 18 provides the correlations for the 4th-grade reading items. Differences of up to .09 were observed between nondisabled students and disabled students with accommodations, and differences of up to .13 were observed between nondisabled students and disabled students without accommodations. Correlations were typically, but not always, lower for the disabled than for the nondisabled students. As expected given the lower mean scores of disabled students, the items that discriminated better for disabled than for nondisabled students tended to be easier items.

There is no clear explanation for the difference between the OR and MC items in the consistency of correlations across groups. The results suggest that the MC items, on average, may be slightly less discriminating for disabled than for nondisabled students, but there were many exceptions to this, and no obvious patterns. The DIF analyses, described next, provide another way of looking at item-level differences between formats.

Table 18

Correlations Between Multiple-Choice Common Item Performance and Theta Scores, Grade 4 Reading

	Item							
	1	2	3	4	5	6	7	8
Nondisabled	.24	.32	.49	.48	.51	.58	.56	.37
SWD, unaccommodated	.26	.32	.53	.51	.50	.58	.55	.38
SWD, accommodated	.26	.19	.50	.45	.40	.50	.55	.36
	Item continued							
	9	10	11	12	13	14	15	16
Nondisabled	.52	.51	.47	.63	.50	.58	.48	.56
SWD, unaccommodated	.55	.54	.40	.56	.54	.52	.43	.57
SWD, accommodated	.49	.53	.46	.51	.49	.46	.41	.53

Note. SWD = students with disabilities.

Differential Item Functioning

A wide variety of procedures have been developed to examine whether an individual item functions differently for members of two groups who have the same total score on a test. If an item is more difficult for one group than for another, when groups are matched on total score, it is said to exhibit *differential item functioning*, or DIF. Items that are flagged as showing DIF may be thought of as measuring a construct that is unrelated to the focus of the test as a whole. Sometimes this is apparent from inspection of the item content. For example, an item on a general test of math achievement that requires students to apply their knowledge of baseball rules might show DIF in favor of males. Often, however, there is no clear source of DIF. Several efforts have been made to identify features of items that commonly exhibit DIF in favor of certain groups (O'Neill & McPeek, 1993; Scheuneman & Gerritz, 1990), but the interpretation of results from a DIF study remains a difficult and often subjective task.

In the 1995 KIRIS data, DIF on the OR items was explored using the logistic discriminant function analysis (LDFA) procedure described by Miller and Spray (1993). Plots of the observed mean item scores for students in different groups at each level of total test score were also examined to provide evidence concerning the magnitude of the DIF.

In the 1995 data, several items on the reading test showed statistically significant DIF, but inspection of the corresponding plots revealed that in nearly all cases the DIF was small enough to have no more than a trivial effect on scores (Koretz, 1997). Six math items, in contrast, showed DIF that appeared substantial upon visual inspection. Some items favored nondisabled students in comparison with accommodated students with disabilities, whereas others favored the latter group over the former. Examination of the item content suggested that items that required extensive reading but that did not require using nonverbal representations such as graphs or tables were differentially easy for accommodated students with disabilities. This seems sensible in the light of the specific accommodations recorded, many of which focused on reading and comprehension. Because of the small number of items involved, however, this explanation is tentative.

We had three objectives in conducting the DIF analyses for the 1997 data. First, we hoped to gather additional evidence to support or refute the hypothesis

concerning verbal load of math items, described above. Second, we wanted to explore DIF in science and social studies in addition to the two subjects examined in 1995. Finally, the inclusion of MC items provided an opportunity to examine differences in the magnitude and frequency of DIF across formats.

DIF on OR Items

As in 1995, we evaluated DIF on the common OR items using the LDFA procedure, which is suitable for items that have more than two response categories. Separate analyses were conducted for students with disabilities who received accommodations and those who did not. The comparison group in each case was the nondisabled student sample. Statistically significant ($p < .01$) DIF was observed for one or both groups on 38 of the 48 common items, representing 79% of all the items across all grades and subjects. Of these, 27 showed DIF for the accommodated students, 3 for unaccommodated students, and 8 for both groups. Non-uniform DIF, which indicates that the magnitude of DIF varied significantly across levels of the total score scale, was observed for 22 of these items. The large number of statistically significant results is due in part to the sensitivity of the logistic regression procedure and the large sample sizes.[4] It is therefore important to supplement these results with the visual inspection to determine which items show DIF that is substantively important.

Figure 10 provides an example of the plots we created for each item to evaluate the magnitude of DIF. The mean item score was plotted for each group of students at intervals along the theta scale of 0.5 standard deviation units. Cells with fewer than 50 students were not plotted. These plots reveal differences in mean scores obtained by each group at various levels of total score and enable us to identify the total score regions where these differences were largest. We consider a group difference on an OR item score of 0.2 or larger (on the 5-point scale) along any region of the theta scale to be moderately large, though this number is arbitrary.

---

[4] We did not make any adjustment for conducting multiple tests. A simple Bonferroni adjustment would have indicated that an individual critical value of .0001 should be used to produce an overall test for the full set of 96 tests at $p = .01$. This would have resulted in 29 items being flagged. We decided not to make this adjustment because the statistical analysis was essentially a screening step that would suggest items whose plots might be worth examining.

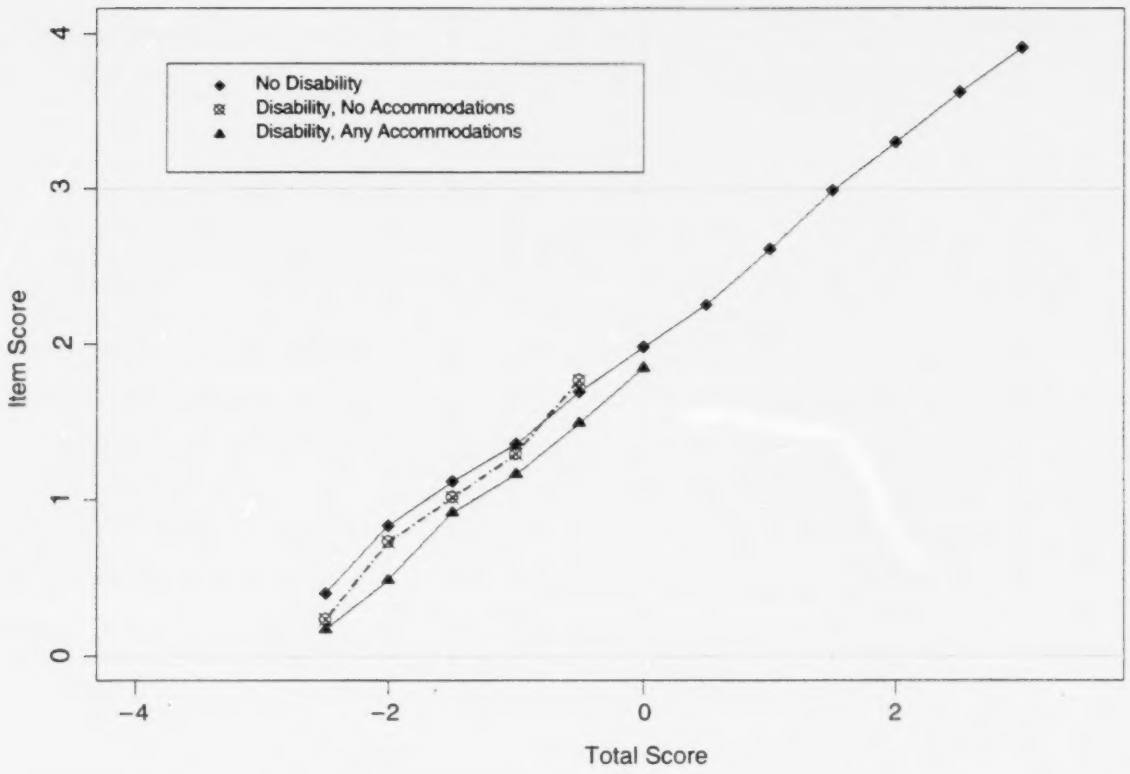


Figure 10. Grade 11, social studies, question 1.

Aside from the math items, only three OR items exhibited DIF that was even moderately large in any region of the theta scale. Two appeared on the 11th-grade test. These included one social studies and one science item, both of which favored nondisabled students over accommodated students with disabilities. The group difference ranged from approximately 0.1 to 0.3 for the region below the mean in the nondisabled population. Because of the sparseness of data for disabled students above the mean, it was not feasible to examine the magnitude of DIF in that region. Figure 10 shows the plot for the social studies item. One item from the 4th-grade reading test exhibited a similar degree of DIF, also favoring nondisabled students over accommodated students with disabilities.

The remaining seven OR items showing moderate DIF appeared on the math tests. These included all four common items at the 5th grade, two at the 8th grade, and one at the 11th grade. For all items the largest difference was observed between nondisabled students and disabled students who received

accommodations, and in most cases the unaccommodated students with disabilities performed similarly to nondisabled students at the same score levels. Two 5th-grade items and both 8th-grade items favored nondisabled students throughout most of the total score range, whereas one 5th-grade item and the 11th-grade item favored accommodated students with disabilities. The remaining 5th-grade item favored students with disabilities below the mean and nondisabled students above the mean. None of the differences was as large as those observed in the 1995 data. An example of one of the larger differences can be seen in the plot depicted in Figure 11. On this item, from the 5th-grade math test, nondisabled students outperformed accommodated disabled students at the same total score points, particularly in the region below the mean in the nondisabled population. Another item from the same test is plotted in Figure 12. This item also favored nondisabled students, but most of the differences were observed above the mean. Figure 13 shows a third item from the 5th-grade test,

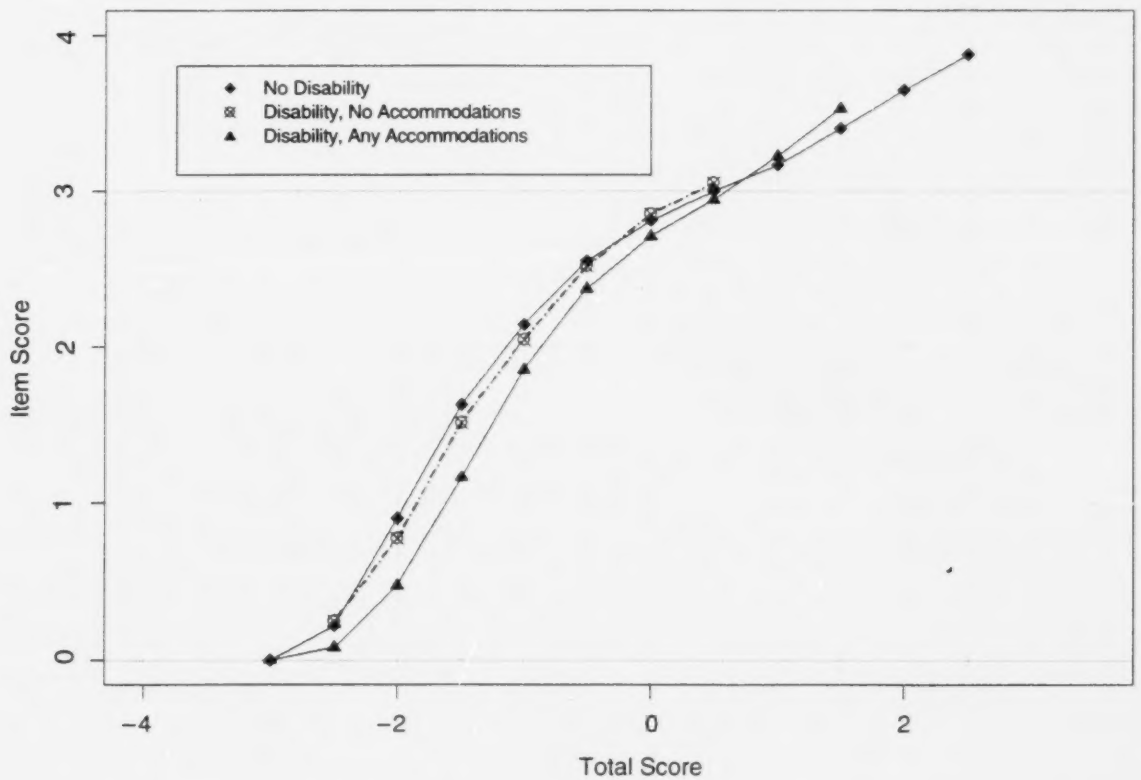


Figure 11. Open response, Grade 5, mathematics, question 4.

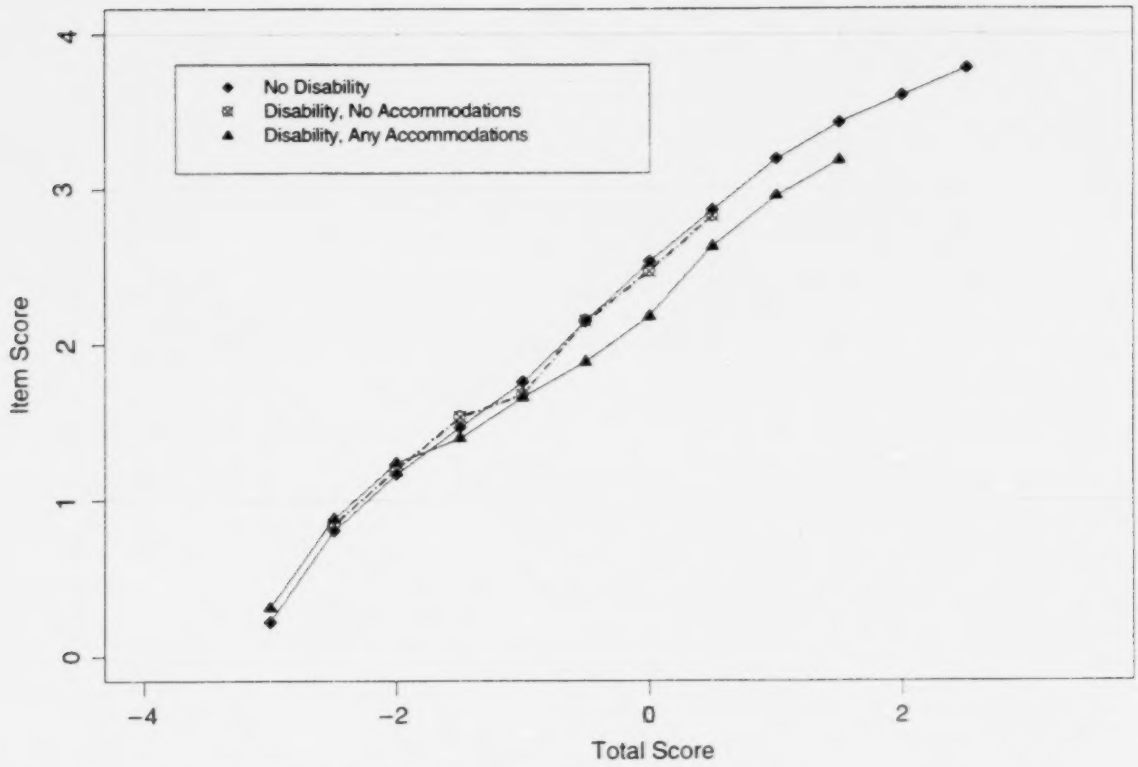


Figure 12. Open response, Grade 5, mathematics, question 2.

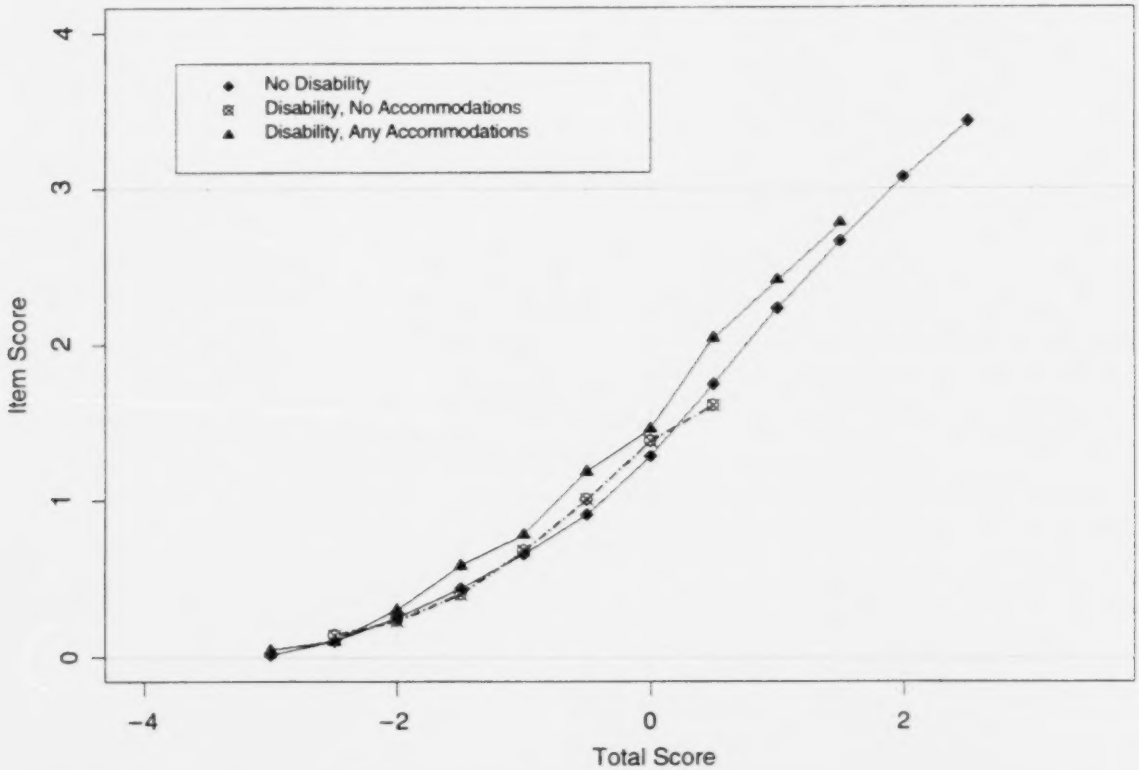


Figure 13. Open response, Grade 5, mathematics, question 3.

this one favoring students with disabilities who received accommodations. These performance gaps were smaller than the largest differences observed in the 1995 data, where groups differed by 0.5 or more on some math items.

Although most of these differences are fairly small, it is worth examining the item content to identify common features that might explain the results. We were especially interested in evidence related to the hypothesis offered by Koretz (1997) regarding the reading and visual processing requirements of the 1995 math items. Examination of the items produced mixed results. The two OR items that favored accommodated students throughout most of the score range appeared to have fairly significant verbal requirements. The 5th-grade item asked students to write a story problem that corresponded to some simple arithmetic equations. Although there was very little reading, the writing requirement probably presented a significant challenge to students who were not comfortable writing. The other item, from the 11th-grade test, involved extensive reading and some fairly advanced mathematical vocabulary (e.g., "depreciation"), so it is conceivable that having some adult assistance on this item was particularly helpful.

Of the four OR items favoring nondisabled students, two had substantial visual content. One involved graph interpretation and the other asked students to identify right angles from a set of diagrams. The remaining two items included a set of series completion problems and a problem evaluating students' understanding of order of arithmetic operations. Examination of the item content as well as the scoring rubrics revealed no clear similarities among these items. Consistent with the hypothesis advanced in our earlier study, however, these items would appear to focus less on verbal demands that would be particularly sensitive to the types of accommodations recorded in Kentucky.

It is important to keep in mind that the DIF analysis identifies items on which some students perform especially well or poorly, given their performance on the test as a whole. Given that DIF was relatively minor for students with disabilities without accommodations, DIF may be interpreted as evidence that accommodations affect performance on some items more than others. Assuming accommodations help disabled students to perform better, items for which DIF favors students with accommodations may be those that are relatively susceptible to the effects of accommodations, whereas those items for which DIF

favors nondisabled students may be those that are relatively resistant to the effects of accommodations. It would be useful for test developers to understand the kinds of items on which performance may be especially resistant to the effects of accommodations and those that may be unusually susceptible to effects that threaten the validity of the items for some students. However, because there are only a few items per grade, and the effects of accommodations vary across grade levels, it is difficult to arrive at any clear conclusion from these data regarding features of items that exhibit DIF.

DIF on MC Items

Several well-tested methods exist for evaluating DIF on MC items. To facilitate consistent interpretation between the OR and MC results, we used a logistic regression procedure. Normally, this procedure involves predicting the item score from group membership, total test score, and their interaction (Swaminathan & Rogers, 1990). However, because our OR DIF detection approach involved predicting group membership from item score, test score, and their interaction, we used the same approach for the MC items. As with the OR DIF analysis, the logistic regression procedure was used primarily to identify items that appeared to warrant a closer look, and was supplemented by plots of the probabilities of a correct response for each group at each total score level. One set of analyses compared nondisabled students with disabled students who received accommodations, and the other set compared nondisabled students with disabled students who did not receive accommodations.

There were a total of 190 common MC items across all grades and subjects (16 for each subject, with four subjects in 11th grade and two in all other grades; 2 items were eliminated by Kentucky because they functioned poorly). Of these, 112 were identified as exhibiting significant DIF for one or more groups. This number represents 59% of the total set of items, a smaller proportion than was identified on the OR test. Table 19 indicates, for each grade and subject, the number of items showing DIF for the accommodated students with disabilities, the number showing DIF for unaccommodated students with disabilities, and the number showing DIF for both of these groups. In all cases, the DIF may be in either direction (i.e., in favor of either nondisabled students or the relevant group of students with disabilities) or may favor one group in some score regions and another group in other regions.

Table 19

Items Identified as Exhibiting Statistically Significant ($p < .01$) DIF

	Number of items		
	Accommodated	Unaccommodated	Both groups
Reading			
Grade 4	10	0	3
Grade 7	11	0	3
Grade 11	2	0	2
Math			
Grade 5	8	0	2
Grade 8	8	1	2
Grade 11	4	0	3
Science			
Grade 4	11	0	2
Grade 7	3	1	4
Grade 11	2	1	5
Social Studies			
Grade 5	7	0	1
Grade 8	4	0	3
Grade 11	3	2	4

For the majority of these items, DIF was observed for the comparison between accommodated students with disabilities and nondisabled students. Some items showed DIF for both accommodated and unaccommodated students with disabilities, and a few showed DIF for unaccommodated students only. The numbers of items exhibiting DIF for accommodated students decreased with grade level, as would be expected by the decreasing use of accommodations in the higher grades. Reading at Grades 4 and 7 had the largest number of items showing DIF; out of 16 items at each grade, 13 showed DIF in Grade 4 and 14 in Grade 7. In contrast to the OR results, DIF was no more frequent for math than for any other subject.

As with the OR items, visual inspection of plots was used to identify items for which the DIF appeared to be practically significant and to determine which group was favored in each part of the score range. The proportion of students answering the item correctly was plotted for each group at intervals of 0.5 along the theta scale. While many of the items showing significant DIF had nearly overlapping curves for the three groups, 42 items showed a difference of at least

0.1 in the proportions of correct responses given by nondisabled students and at least one group of disabled students at two or more plotted points along the score range. In most of these cases, nondisabled students were favored over disabled students who received accommodations, but there were some items that favored disabled students. Figure 14 shows one example of the latter type. As was typical for most items, the differences were largest in the region just below the mean in the nondisabled population, a region that included a substantial proportion of the students with disabilities. This item, which asked students to select an explanation for pitch change in a tightened guitar string, clearly favored accommodated students over both other groups in this region. Inspection of the item content revealed no clear trend in the kinds of multiple-choice items likely to exhibit DIF.

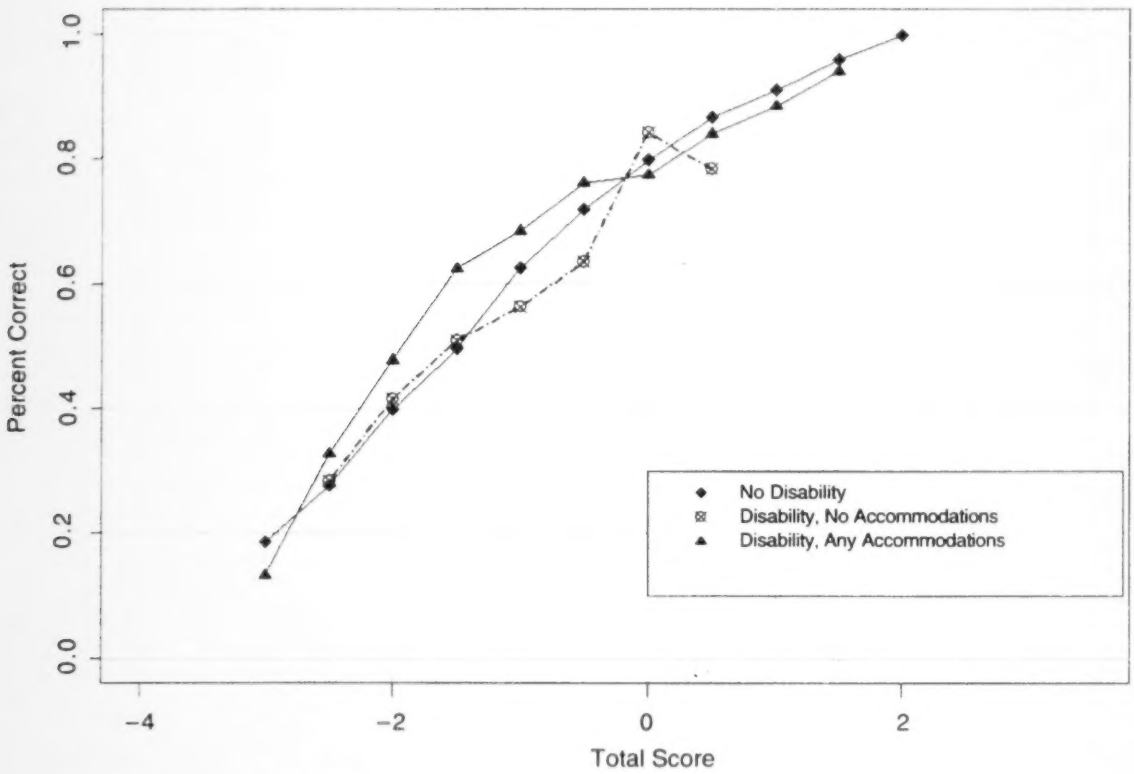


Figure 14. Multiple choice, Grade 4, science, question 4.

Comparison Between Formats

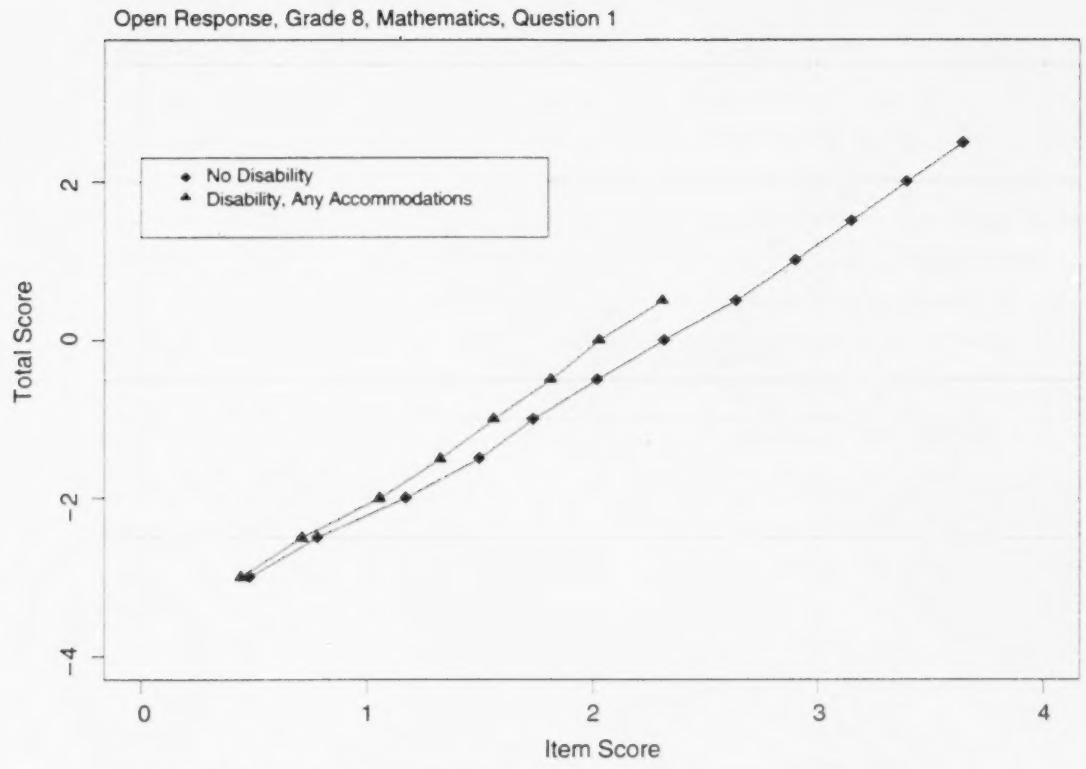
As mentioned earlier, one of the primary reasons for conducting this set of DIF analyses was to explore differences in the frequency and magnitude of DIF between the OR and MC formats. Because the two formats use different scales, the analyses and plots presented above do not facilitate such a comparison. To provide a way of comparing DIF between the formats visually, we reversed the axes in the plots used in the previous section, so that average total score is plotted as a function of item score (mean item score on the 5-point scale for OR items, or percent correct for MC items). Thus, the vertical distance between groups on the y axis in these plots has the same interpretation for both formats: the mean difference in total scores, in standard deviation units (standardized in the nondisabled population), between groups of students whose total scores predict a given, equivalent level of performance on a specific item. We selected the items in each format that showed the largest DIF for a given grade and subject, and compared the magnitudes of the total score differences across levels of item score.

The largest difference in total score on the OR items between groups at any level of item score was approximately one half of a standard deviation unit. Differences of this magnitude were observed on seven items, which represent approximately 15% of all the OR items. Five of the items showing this degree of DIF were math items. A number of MC items also showed differences this large, but unlike on the OR test, these were spread across all subjects. Twenty-six MC items, representing approximately 14% of the total, exhibited this degree of DIF. In addition, several MC items showed differences substantially larger. Differences of approximately one standard deviation unit were observed on 8 items (4% of the total), and differences of one and one half standard deviation units were observed on 3 additional items (1.5% of the total). Nine of these 11 items showing especially large differences favored nondisabled students over disabled students who received accommodations.

Figure 15 shows plots for the OR and MC 8th-grade math items showing the largest degree of DIF.[5]

---

[5] Note that because the axes are transposed, the curve for the group that is favored by the item lies *below* the curve for the other group.



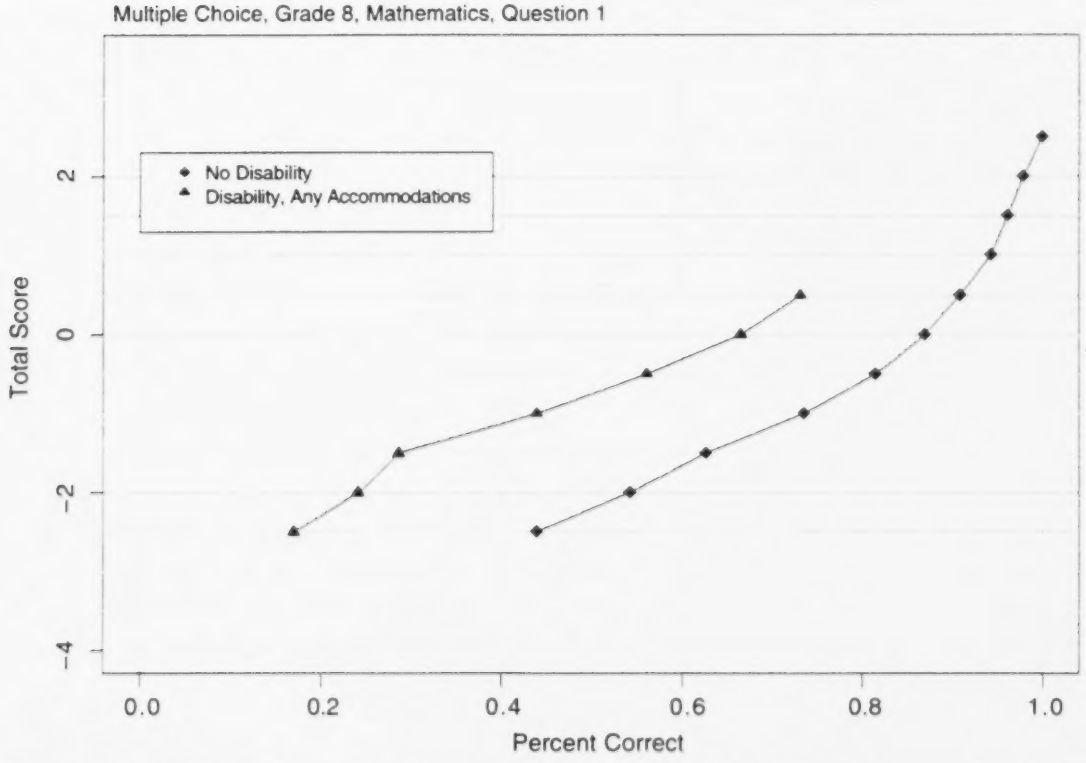


Figure 15. Plots for open-response and multiple-choice 8th-grade math items showing the largest degree of DIF. Note that because the axes are transposed, the curve for the group that is favored by the item lies *below* the curve for the other group.

The degree of DIF observed on the MC item was substantially larger than that on any of the OR items. For the MC item, at every point on the percent correct scale where there were sufficient numbers of students to plot points for both groups, the difference between total scores was at least 1.2 standard deviation units. The largest difference observed for the OR item, in contrast, was less than .5 standard deviation units.

In 1995, several OR items showed DIF larger than any DIF that was observed in the 1997 OR items. Thus, we see improvement in the functioning of the 1997 test in one respect: The OR items seem to function more equivalently for disabled and nondisabled students than they did in 1995. However, this improvement is offset to some degree by the relatively large DIF that we observed on several of the MC items.

Discussion

In an earlier study that looked at Kentucky's assessment of students with disabilities in 1995 (Koretz, 1997), we concluded that Kentucky had successfully included the great majority of students with disabilities in its regular statewide assessment. We concluded, however, that there appeared to be serious weaknesses in the assessment of these students, including an apparent overuse of certain accommodations, implausibly high scores for students who received certain accommodations, excessive difficulty levels for students with disabilities in mathematics, and substantial DIF in mathematics for disabled students assessed with accommodations.

The present study extended the earlier work in three ways: It examined the consistency of these findings two years later; it extended some of the analyses into additional grades and subjects; and it included multiple-choice as well as open-response test items. The findings of the present study are in some respects more optimistic in terms of the quality of measurement, although they are more pessimistic about the performance of certain groups of students with disabilities. They also show potentially important interactions among disability, accommodations, and test format that need further investigation.

Because of the complexity of the results, this section begins with a summary of important findings. This is followed by a discussion of some implications of the findings for policy and research.

Summary of Findings

The reported use of accommodations remained largely unchanged from 1995 to 1997. Most students with disabilities received accommodations, and the majority of those who were given accommodations were provided with more than one. The use of accommodations was most extensive in the elementary grades and was lowest in the 11th grade. As in 1995, teachers in 1997 reported making extensive use of accommodations that one would expect to see used infrequently. For example, roughly half of all elementary and middle school students with disabilities had at least part of the test paraphrased, despite state guidelines that indicate that paraphrasing should be used only sparingly. The similarity of these findings to those in 1995 is striking, and the findings are important for two reasons. First, they confirm that the seemingly problematic uses of accommodations seen in 1995 were not anomalous. Second, the similarities remove one of the most obvious possible explanations for the considerable declines from 1995 to 1997 in the performance of some students with disabilities—that is, the possibility that these drops were caused by a change in the frequency with which disabled students were provided with accommodations.

In 1997, the mean scores of all students with disabilities, aggregated without respect to specific disabilities or accommodations, were substantially lower than those of students without disabilities. This was true in all grades and subjects and on both the open-response and multiple-choice components of the assessment. However, the gap was much larger among older students; it ranged from 0.4–0.7 standard deviation in the 4th grade to 1.0–1.4 standard deviations in the 11th grade. Differences in performance across the two test formats were modest but differed by grade. The gap between disabled and nondisabled students tended to be bigger on the multiple-choice components in the elementary grades, whereas the gap was generally similar or larger for open-response components in the higher grades.

These findings represent a sharp drop in the mean performance of elementary school students with disabilities. This decline stemmed from the performance of students who received accommodations; the mean performance of disabled students without accommodations remained quite stable. Examination of the performance of students with learning disabilities or mild

mental retardation—the only groups large enough for this specific analysis—showed a sizable drop in the mean scores of students who received dictation in combination with oral presentation, with or without paraphrasing. These are precisely the groups of disabled students whose scores we deemed implausibly high in 1995. For example, the average scores of learning-disabled fourth-grade students who received these accommodations were as much as 0.5 standard deviation above the mean for nondisabled students in 1995. In contrast, in 1997, the mean scores of these students in the fourth and fifth grades ranged from slightly above the average of nondisabled students to 0.5 standard deviation below, and they were below the mean of nondisabled students in six of eight cases. These scores appear on their face more plausible that the corresponding, higher scores from two years earlier.

Regression analysis, used to disentangle the associations between performance and accommodations that were usually offered in combinations, confirmed that a change in the correlates of dictation underlay much of the drop in the performance of elementary students with disabilities on the open-response components of KIRIS. In 1995, the mean estimated independent association between dictation and KIRIS scores among learning-disabled students (across all grades and subjects) was 0.7. That is, learning-disabled students receiving dictation scored, on average, fully 0.7 standard deviation above learning-disabled students who did not receive dictation, holding constant other accommodations, grade, and subject. In 1997, the mean association on the open-response components had dropped to 0.4 standard deviation. Because dictation is offered primarily to students in the elementary grades, this change would depress the mean of elementary students with disabilities more than the means of older students, consistent with the patterns we found.

The effects of three accommodations—dictation, paraphrasing, and cueing—tended to be larger for the open-response components of the assessment than for the multiple-choice components. This is not surprising in the case of dictation, because this accommodation would seem not to be relevant on the multiple-choice components for most students. Cueing could be relevant on both components, but it would seem most relevant to open-response items that require complex student responses. In contrast, oral presentation did not show a consistent format difference, but there was only a single instance in which oral presentation showed more than a trivial positive association with scores.

The correlations among parts of the assessment provide hints about the extent to which scores represent the same dimensions of performance for disabled and nondisabled students. A high correlation among parts suggests that the parts measure a common dimension or aspect of performance, while lower correlations suggest that one part is substantially influenced by something not tapped by the other part. Differences in these correlations between disabled and nondisabled students were generally not large, but some were substantial enough to warrant note, and the patterns they showed are suggestive.

First, correlations between parts within subject areas tended to be lower for disabled students with accommodations, particularly in the higher grades. This was not true of disabled students without accommodations. This suggests that accommodations (or other characteristics of students who received accommodations) caused scores on one or both parts to be influenced by factors that had less impact on the scores of other disabled and nondisabled students. These factors may be irrelevant to the constructs the assessment is intended to measure.

Second, for students with disabilities who received accommodations, many of the correlations among open-response scores in different subjects were larger than the correlations among scores in the same subject areas but across formats. These correlations were substantial for all students, which suggests that a single dimension of performance, presumably entailing skills in reading and writing, has a substantial impact on scores on the open-response components, regardless of subject. These correlations suggest that this dimension influences open-response scores of accommodated students more than does subject-specific knowledge.

Correlations between scores and performance on items show the extent to which items differentiate among high- and low-performing students and provide another view of the homogeneity or dimensionality of the assessment. Consistent with the 1995 Kentucky data—but inconsistent with some other studies—the 1997 data showed no substantial differences in these correlations among nondisabled students, disabled students without accommodations, and disabled students with accommodations. In contrast, in the case of the multiple-choice components, correlations between scores and item performance differed across groups, with many correlations lower for disabled students than for nondisabled students. There were numerous exceptions, and some items were

more discriminating for one of the disabled student groups. These variations in discrimination would not be surprising if easier items discriminated better and hard items discriminated more poorly for students with disabilities. Some items did not conform to this pattern, however, and further research would be needed to explain these findings.

In the 1995 data, we found important instances of differential item functioning (DIF), but they were largely limited to mathematics (only mathematics and reading were analyzed) and to students receiving accommodations. Some items favored accommodated students with disabilities, whereas others favored nondisabled students at the same score level. In the 1997 open-response data, DIF again affected mostly mathematics (in this case, four subjects were analyzed) and students with accommodations. The largest instances of DIF were smaller than those in 1995, perhaps another example of more accurate assessment of students with disabilities in 1997. The multiple-choice data from the 1997 assessment, however, presented a very different picture. On these components, a smaller proportion showed DIF, but DIF was apparent in all subject areas. Indeed, in contrast to the open-response components, DIF on the multiple-choice components was most common in reading. DIF on multiple-choice items generally favored nondisabled students. In addition, the largest instances of DIF on multiple-choice items were far larger than the largest on open-response items.

Implications

The substantial decline in the mean performance of elementary school students with disabilities in Kentucky ironically may be good news in terms of measurement. In our earlier study (Koretz, 1997), we stressed the implausibly high mean scores of some groups of students with disabilities as an indication that the quality of measurement for some students with disabilities was poor. The lower mean scores observed two years later are more plausible on their face, and there is ancillary evidence further suggesting that they are more reasonable. The independent associations between performance and accommodations (particularly dictation) were far more modest in 1997, and DIF on open-response items was ameliorated somewhat from two years earlier. These findings are, of course, insufficient to demonstrate the validity of scores for students with

disabilities, but they do eliminate some of the more patent indications of invalidity.

The comparisons between open-response and multiple-choice components of KIRIS have important implications and raise significant questions. Some observers have argued that performance assessments will pose less of a barrier to many students who do poorly on traditional, multiple-choice assessments, while others have argued that assessments that require extensive reading and writing will further disadvantage some students with disabilities, particularly those with learning disabilities. Neither appears to be clearly true in the case of the KIRIS assessment. While there were some appreciable differences in mean scores across the two formats, many scores were similar, and neither format consistently favored students with disabilities across all ages and subjects. The DIF analysis also did not suggest strongly that one of the formats is consistently better for students with disabilities. While the largest instances of DIF occurred in the multiple-choice component, sizable DIF was a bit less common in that component.

These findings about format differences, however, may depend on both the particular attributes of KIRIS and the current uses of accommodations in that assessment. Under other conditions—for example, if the uses of accommodations were subject to different constraints than those currently in place in Kentucky, or if the difficulty level of the assessment were different—the impact of format could be quite different. It is important to bear in mind that Kentucky has no data on the uses of accommodations in the various components of the assessment. It is therefore entirely possible that accommodations were offered differently on the open-response and multiple-choice components of the assessment and that this contributed to the relative similarity of scores.

In our earlier study, we presented speculation that the impact of accommodations stemmed in part from their effects on the verbal difficulty (both reading and writing) of open-response items. There are further suggestions in the present findings that verbal difficulty plays a different role in the assessment of accommodated students with disabilities than in the assessment of others, perhaps because of the effects of accommodations. One indication is that DIF was distributed quite differently across subjects in the open-response and multiple-choice components of the assessment. On the open-response components, DIF

was least common in reading (consistent with the results two years earlier). In contrast, on the multiple-choice component, DIF was *most* common in reading. Another indication was provided by the correlations among parts of the assessment. In the case of accommodated students with disabilities, the correlations across subjects of open-response scores were often larger than the correlations between scores within subjects but across formats. This suggests that the reading and writing demands of the open-response components influence the performance of accommodated students with disabilities more than subject-matter knowledge.

There is a clear need for additional descriptive studies of the performance of students with disabilities in large-scale assessments. In our earlier study, we noted that research evidence was sparse and argued that "descriptive studies similar to this one but in different settings and with different assessments are needed to explore the generalizability of the findings" (Koretz, 1997, p. 67). The present findings make this need even more apparent. The large changes in performance and its correlates observed in Kentucky from 1995 to 1997 impede generalization to other assessments. Similarly, the findings of the present study pertaining to format differences raise as many questions as they answer.

Descriptive studies can only be as good as the data on which they are based, and the results of the present study also underscore the need for strengthened collection of data pertaining to students with disabilities participating in large-scale assessments. Other jurisdictions are experimenting with collecting simple descriptive data on disabilities and accommodations in their assessments. If it is overly burdensome to obtain this information with each assessment, it could be collected periodically, say, twice per decade. More detailed data on the use of accommodations—for example, an indication of the specific portions of the assessment for which they are offered—would allow more informative analysis.

In our earlier study, we pointed out the limitations of non-experimental analysis (including our own) for determining the effects of accommodations and called for true experimental studies of accommodations. Some experimental studies of accommodations have indeed been undertaken recently (e.g., Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). We reiterate here the call for additional experimental studies, but it is important to recognize that there may be serious constraints on the feasibility of experimental trials in the context of large-scale assessments. Legal constraints may preclude experimental trials in the

context of operational assessments, at least when assessment results have consequences. Advocates or policymakers in some jurisdictions may also find a no-accommodations condition unacceptable, even in field trials. If that were to happen, experiments would be limited to comparisons among potentially acceptable types of accommodations. Finally, the number of combinations of assessments and disabilities that could be handled by a given experiment is likely to be small, particularly in the light of the relatively small numbers of students in each group and the implausibility of fractional designs that assume a lack of interactions between disabilities and accommodations.

Accordingly, it may be important also to carry out more detailed non-experimental studies that would require richer data than can be obtained through routine data collection. It might be feasible to obtain the needed data by piggybacking additional data collection on the administration of ongoing large-scale assessments. For example, teacher surveys could obtain additional information about the characteristics of students, the accommodations offered to them in instruction and on other assessments, and their performance on different measures of achievement. These data would provide a much more complete descriptive view of assessment and accommodations and a stronger basis for hypothesizing about the effects of format, accommodations, and other factors.

Even with an increase in research, it will be some years before the field can provide policymakers and practitioners with strong guidance about the assessment of students with disabilities and the use of accommodations. In the interim, however, policymakers can monitor the assessment of students with disabilities. For example, routine data collection should be sufficient to uncover some of the problems noted here, such as the apparent overuse of accommodations and the excessive difficulty of parts of the assessment for students with disabilities. Finally, policymakers can make use of monitoring data and the slowly emerging research literature to refine assessment guidelines, inform test development, and guide the development of policy.

References

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994.* Frankfort, KY: Office of Education Accountability, Kentucky General Assembly, June.

Kentucky Department of Education. (1996). *Procedures for inclusion of all students in the KIRIS accountability and state-required national-reference assessments* [Draft]. Frankfort, KY: Author.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431). Los Angeles: University of California, Center for the Study of Evaluation.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

National Research Council, Committee on Goals 2000 and the Inclusion of Students with Disabilities. (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

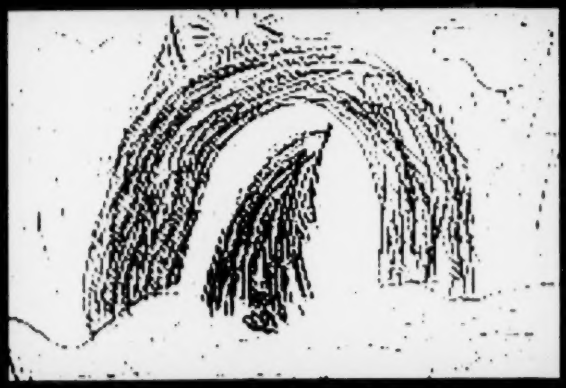
O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

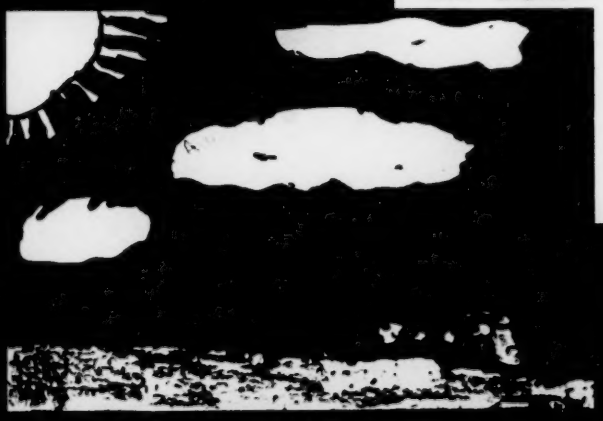
Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27,* 109-131.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64,* 439-450.

U.S. Department of Education. (1996). *To assure the free appropriate public education of all children with disabilities: Eighteenth annual report to Congress on the implementation of the Individuals With Disabilities Education Act.* Washington, DC: Author.





UCLA Graduate School of Education & Information Studies